A Simple App to Teach Regression, http://trialandstderror.com/ *

Luke M. Froeb†

30 December 2020

Abstract

This paper introduces a web app for teaching regression that "inverts" the usual pedagogy. Rather than asking students to run regressions on data, it asks them to "create" data to achieve a given outcome, like a statistically significant line. By clicking on an (x, y) graph, students watch as the regression line, with its confidence intervals and statistics, is redrawn with each new data point. We describe three self-guided exercises, designed to teach the ideas of statistical significance, and correlation vs. causality.

JEL classification: A2 (Economic Education)

Keywords: Teaching Regression; Teaching Multiple Regression; Dummy Variable Regression, Differences Between Groups, Within Estimators.

*http://trialandstderror.com/, The program was originally written in Hypercard in 1988 to teach Justice Dept. attorneys enough about regression to allow them to cross-examine rival experts. Keyuan Jiang (kjiang@pnw.edu) then ported it over to Visual Basic and then to a javascript, which dramatically improved its look and feel. Taylor Jones (tjonesster@gmail.com) and Gray Curtis (curtisg@alum.mit.edu) updated the web app. Michael Ward gave us helpful comments.

†Vanderbilt University, Owen Graduate School of Management, 401 21st Avenue South, Nashville, TN 37203, USA. e-mail: luke.froeb@vanderbilt.edu

1 Introduction

Although there are exceptions, like Joshua Angrist's "Mastering Metrics" course at Marginal Revolution University[1], regression is traditionally taught with a a lot of analytic overhead. In many classes, students might have to learn random variables, probability, statistics, and hypothesis testing before getting to regression This can put the methodology out of reach to those who could benefit most from using it. This paper introduces a simple web app for teaching regression, accessible by anyone who can point and click.

The app "inverts" the usual pedagogy. Rather than teaching students how to run regressions on data, it asks them to create data to achieve a specified result, like a statistically significant line. Exercises are designed to give students an intuitive feel for the relationship between data and regression, and to show them how regression is used.

The app has self-guided exercises, but can also be used in a lecture or as homework. It may be particularly useful for classes that use regression, e.g., Managerial Economics or Health Economics, but which do not require econometrics as a prerequisite. It may also be useful in econometrics classes to illustrate ideas like correlation vs. causality.

In what follows, we introduce the tool by describing three self-guided exercises, designed to teach the ideas of statistical significance, and correlation vs. causality.

2 Statistical Significance

In this section we describe a self-guided exercise to show students how confidence intervals are related to hypothesis testing and t-stats. Students begin by clicking on the "Learning Exercises" button and then on "Line Confidence Interval" which brings up the following question:

[1] https://mru.org/mastering-econometrics

QUESTION: Line Confidence Interval

Check the line confidence interval box, and uncheck the point confidence interval box.

Create a data set with two properties:

1. You cannot draw a horizontal line inside the 95% line confidence interval.
2. You can draw a horizontal line inside the 99% line confidence interval.

Click on the radio buttons 95% or 90% to see the two line confidence intervals.

A student answers the question by trial and error, eventually producing a data set with the two asked-for properties, illustrated by two graphs like those in Figure 1. When the task is done, the student clicks on the "Am I right?" button which indicates a correct answer or not, and brings up the following answer:

ANSWER: Denoted by two asterisks after the slope t-stat, statistical significance at a 95% level means two things:

1. You cannot draw a horizontal line inside the 95% confidence interval. In other words, you reject the hypothesis that the line is flat (no relationship between X and Y).
2. You can draw a horizontal line inside the 99% confidence interval; otherwise, the line would be significant at a 99% level, denoted by three asterisks after the t-stat.

Figure 1: 95% vs. 99% Line Confidence Interval and Statistical Significance

3 Correlation vs. Causality

One can give the next two exercises a concrete context by using the example from Professor Angrist's "Ceteris Paribus" and "Selection Bias" videos at Marginal Revolution University.[2] Let:

Y = post-school wages,

X = student quality, measured by test scores,

Black dots represent private university students, and

Blue dots represent public university students.

In the three exercises that follow, we estimate differences in performance between private university students (black dots), and public university students (blue dots). The pedagogy is to compare and contrast the results of simple mean differences that ignore test scores to regression estimators that take account of them. The lessons are that (i) correlation need not imply causality (false positive); and (ii) lack of correlation need not imply a lack of causality (false negative).

3.1 Between Group Difference (False Positive)

In this section we describe a self-guided exercise to show students that correlation (private university students earn higher wages Y) need not imply a that private university education causes higher wages.

Click on the "Differences Between Groups" button, and then on the "Learning Exercises" button. The first learning exercise brings up the following question:

QUESTION: Mean Difference vs. Dummy Variable I

Create a data set with two properties:

[2] https://mru.org/mastering-econometrics

1. The Mean Differences methodology shows that the two groups are significantly different, denoted by two asterisks after the t-stat; and

2. The Dummy Variable methodology shows no significant difference between the groups.

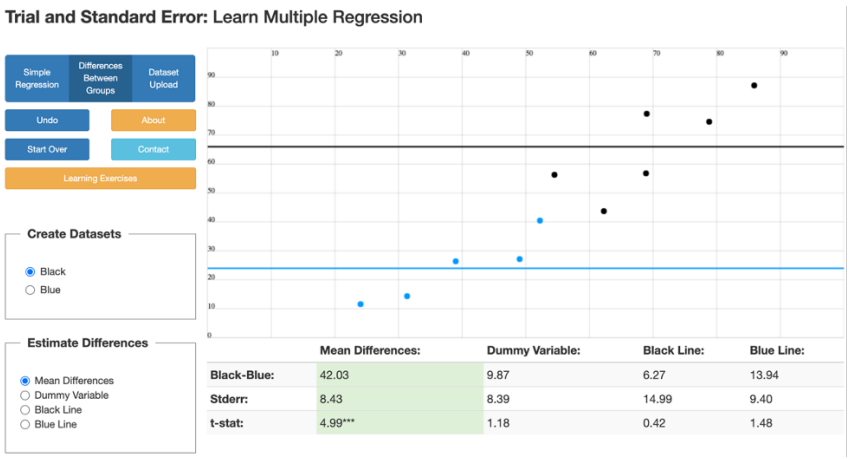
Check your answer by clicking the button "Am I right?," and explain in words what is going on.

A student answers the question by trial and error, eventually producing a data set with the two asked-for properties, illustrated by two graphs in Figure 2. The top graph indicates a significant mean difference while the bottom indicates an insignificant difference from a dummy variable regression including X. When the task is done, the student clicks on the "Am I right?" button which indicates a correct answer or not, and brings up the following answer:

ANSWER: If you look only at mean differences, you can mistakenly infer that private university membership causes higher Y values when, in fact, they are caused by the higher test scores of private university students. This is sometimes referred to as "selection bias" (those with higher test scores values "select" private university more frequently), "omitted variables bias," (estimates of the effect of private university education on wages may be biased biased if you omit test scores from your analysis).

The top panel in Figure 2 shows that the private university group (black) has a mean wage that is much higher than that of public university students (blue). However, this raw difference does not take account of the difference in test scores (X). It is not blue group membership that causes higher wages, rather, it is their higher test scores.

The bottom panel shows a dummy variable estimator of the difference. It shows that the size of the group difference (the vertical distance between the



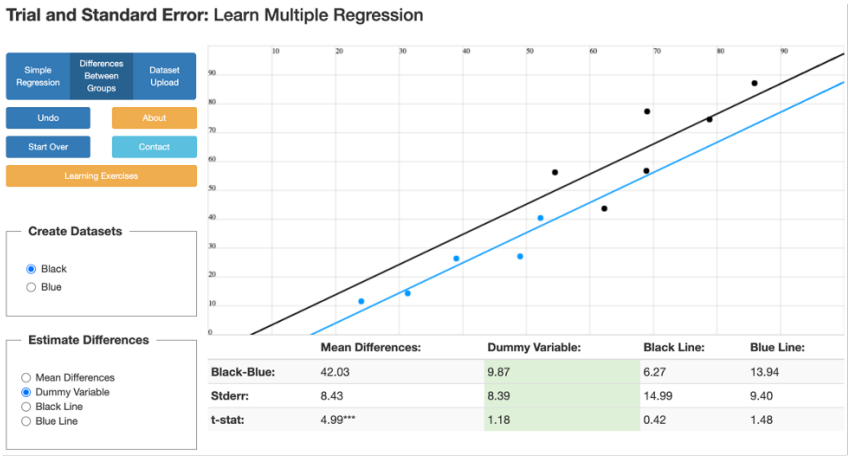


Figure 2: Between Group Difference (False Positive)

black and blue lines) shrinks by more than 50% (and we lose statistical significance), once we account for the difference in X values. Similarly, in Professor Angrist's video on Selection Bias, the mean difference in wages between public and private university students shrinks from 16% to 9% once students' test scores are taken into account.

3.2 Between Group Difference (False Negative)

In this section we describe a self-guided exercise to show students that a lack of correlation (private university students earn same wages as public university students) need not imply that a private university does not cause higher wages.

Click on the "Differences Between Groups" button, and then on the "Learning Exercises" button. The second learning exercise brings up the following question:

> QUESTION: Mean Difference vs. Dummy Variable II
>
> Create a data set with two properties:
>
> 1. The Mean Differences methodology shows that the two groups are not significantly different, denoted by zero asterixes after the slope t-stat; and
> 2. The Dummy Variable methodology shows a significant difference between the groups, denoted by two asterixes after the slope t-stat.
>
> Check your answer by clicking the button "Am I right?," and explain in words what is going on.

A student answers the question by trial and error, eventually producing a data set with the two asked-for properties, illustrated by two graphs like those in Figure 3. When the task is done, the student clicks on the "Am I right?" but-

ton which indicates a correct answer or not, and brings up the following answer:

> ANSWER: If you look only at mean differences, you can mistakenly infer that black group membership does not cause higher Y values when, in fact, it does.

The top panel in Figure 3 shows that the private university group (black) has a mean wage that similar to that of public university students (blue). However, this raw difference does not take account of the difference in test scores (X). In this case, the public university group (blue) has much lower performance than a private university student (black) with similar test scores.

The middle panel shows a dummy variable estimator of the between-group difference, holding test scores constant. It shows that if the public university students (blue) were treated similarly to private university students (black), they would be on the black line, earning much higher wages. Once we account for the difference in test scores, we see a huge causal effect of private university education on wages. The vertical difference between the lines can be thought of as a measure of the effect of a private university education, holding student quality constant.

The third panel illustrates the between-group difference estimated by a regression of wages on student test scores for the black data, and testing whether the blue data are significantly below from the black line, e.g., Froeb et al. (1993). Like dummy variable regression, it shows that if public university students were were treated similarly to private university students, they would be on the black line, earning much higher wages.

4 Conclusion

The benefit of the app is that it gives students an immediate way to "see" what regression does, how it is related to data, and how it is used.

One potential cost is that it shows students how to generate data to determine a result, which may encourage "regression fishing" for desired outcomes.

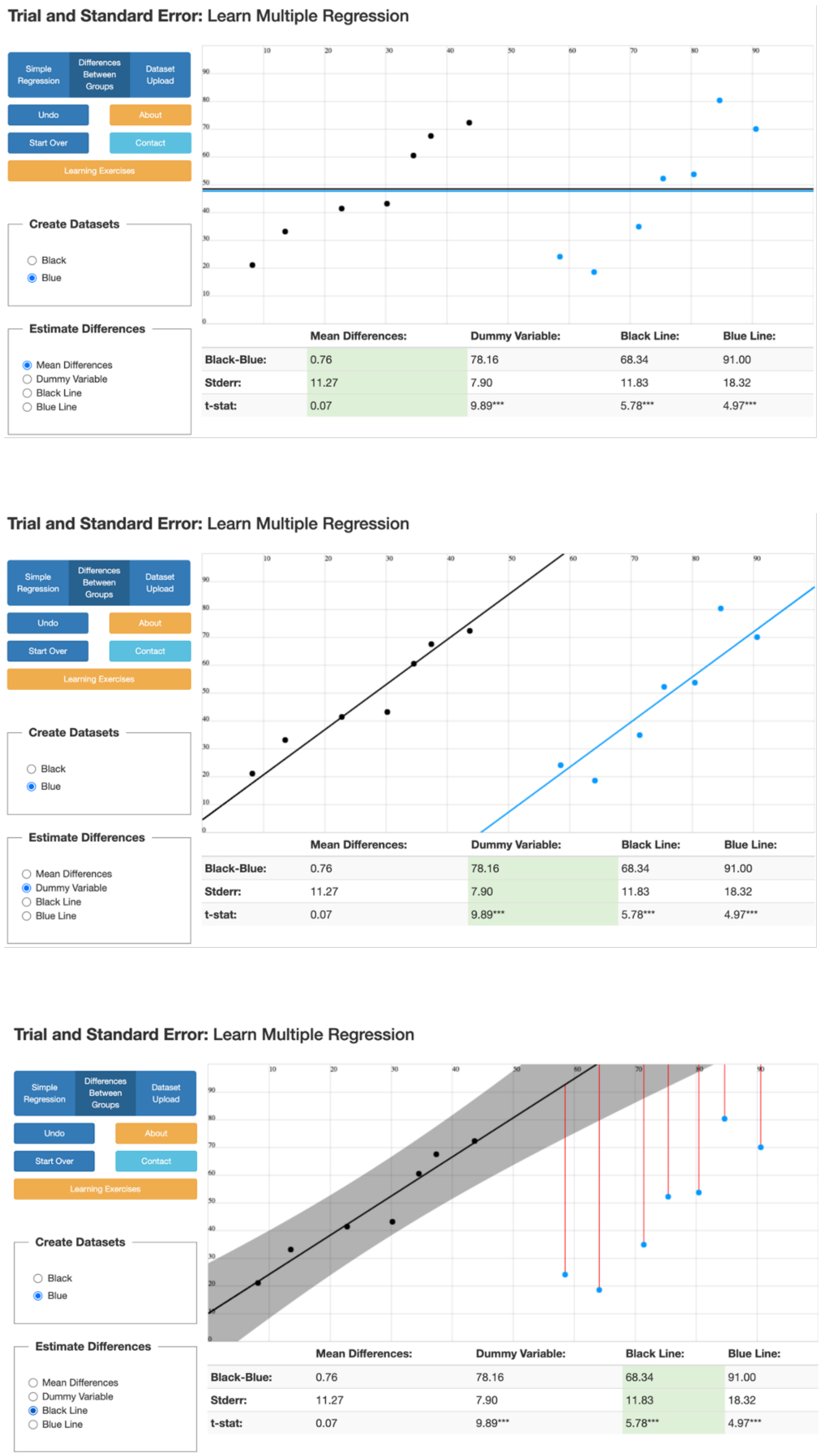


Figure 3: Between Group Difference (False Negative)

However the app may also teach them how to be skeptical of regression results, the bigger goal.

The app is experimental and we hope to get feedback (luke.froeb@vanderbilt.edu). A link to the app is in the first footnote.

References

Luke Froeb, Robert Koyak, and Gregory Werden. What is the effect of bid rigging on prices? *Economics Letters*, 42:419–423, 1993.