Journal of Statistics Education ›

Volume 10, 2002 - Issue 1

Free access

3,582 Views | 1 CrossRef citations to date | 0 Altmetric

Listen

Articles

More on Venn Diagrams for Regression

E. Kennedy Peter

Pages 1-10 | Published online: 01 Dec 2017

Cite this article https://doi.org/10.1080/10691898.2002.11910547

Full Article | Figures & data | References | Citations | Metrics

Reprints & Permissions | View PDF | Share

Abstract

A Venn diagram capable of expositing results relating to bias and variance of coefficient estimates in multiple regression analysis is presented, along with suggestions for how it can be used in teaching. In contrast to similar Venn diagrams used for portraying results associated with the coefficient of determination, its pedagogical value is not compromised in the presence of suppressor variables.

Keywords:

Detrending | Multicollinearity | Omitted regressor | Regression graphics | Teaching statistics

1. Introduction

In a recent article in this journal, Ip (2001) discussed the use of Venn diagrams for

Article contents | Related research

inform readers of several applications of Venn diagrams to regression analysis that were not included in Ip's presentation, and to demonstrate some ways of using these applications to improve student understanding of ordinary least squares (OLS) regression. Because these alternative applications are not new to the literature, the main contribution of this paper consists of suggestions for how this approach can be used effectively in teaching.

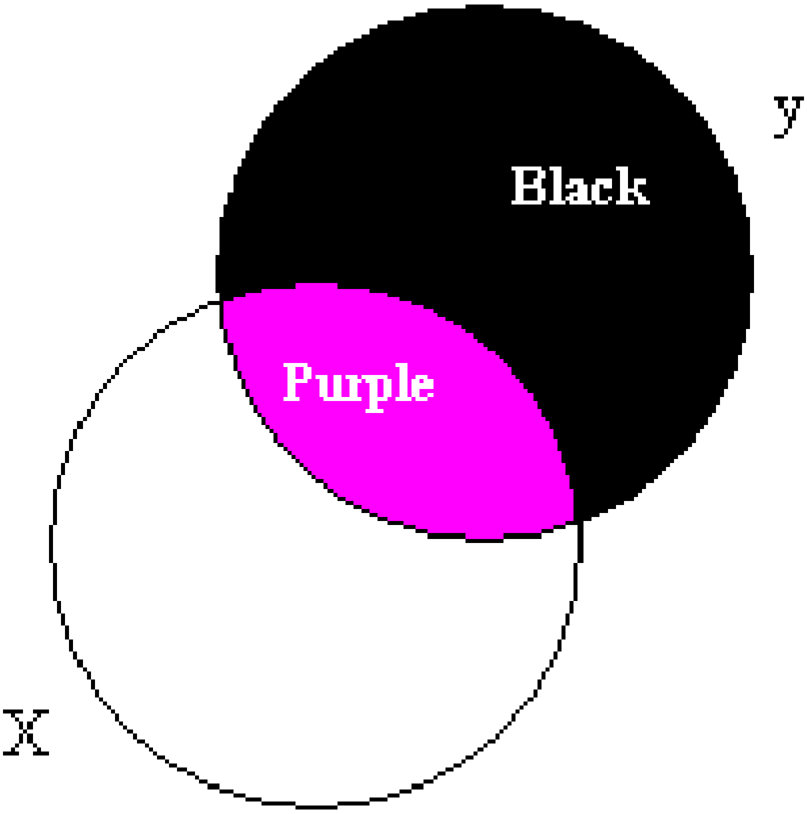
The first use of Venn diagrams in regression analysis appears to be in the textbook by Cohen and Cohen (1975). A major difficulty with its use occurs in the presence of suppressor variables, a problem discussed at length by Ip (2001). No one denies that Venn diagrams can mislead, just as no one denies that ignoring friction in expositions of physical phenomena misleads, or using Euclidian geometry misleads because the surface of the earth is curved. Such drawbacks have to be weighed against the pedagogical benefits of the “misleading” expository device. As recognized by Ip, in the case of applying the Venn diagram to regression analysis, reasonable instructors could disagree on the pedagogical value of the Venn diagram because of the suppressor variable problem.

Ip's article is confined to the use of Venn diagrams for analyzing the coefficient of determination R^2, partial correlation, and sums of squares. In these cases, exposition is compromised in the presence of suppressor variables. But there are other concepts in regression analysis, thought by many to be of considerably more importance than R^2, which are not complicated by suppressor variables, the prime examples being bias and variance of coefficient estimates. This article presents a different interpretation of Venn diagrams, highlighting illustrations of bias and variance, and discusses how these diagrams can be used to enhance the teaching of multiple regression.

2. An Alternative Interpretation

Kennedy (1981) extended the Venn diagram to the exposition of bias and variance in the context of the classical linear regression (CLR) model, written as $y = X\beta + \varepsilon$. Here the dependent variable y is a linear function of several explanatory variables represented by the matrix X, plus an independent and identically distributed error ε. The usual CLR assumptions, such as that the expected value of ε is zero, and that X is fixed in repeated samples, are invoked. For expository purposes, y and X are measured the crucial element is how areas on the diagram are interpreted. In Figure 1 the circle labeled y represents “variation” in y, and the circle labeled X represents “variation” in X, where, for pedagogical purposes, “variation” is not explicitly defined but is left as an intuitive concept. The overlap between the y and X circles, the purple area in Figure 1, is interpreted as “variation” that y and X have in common - in this area y and X “move together.” This co-movement is used by the OLS formula to estimate β_X, the slope coefficient of X.

Figure 1. Venn diagram for regression.



Display full size

Although the purple area represents the variation in y explained by X, just as in Ip's application, Kennedy extends its interpretation in three substantive ways:

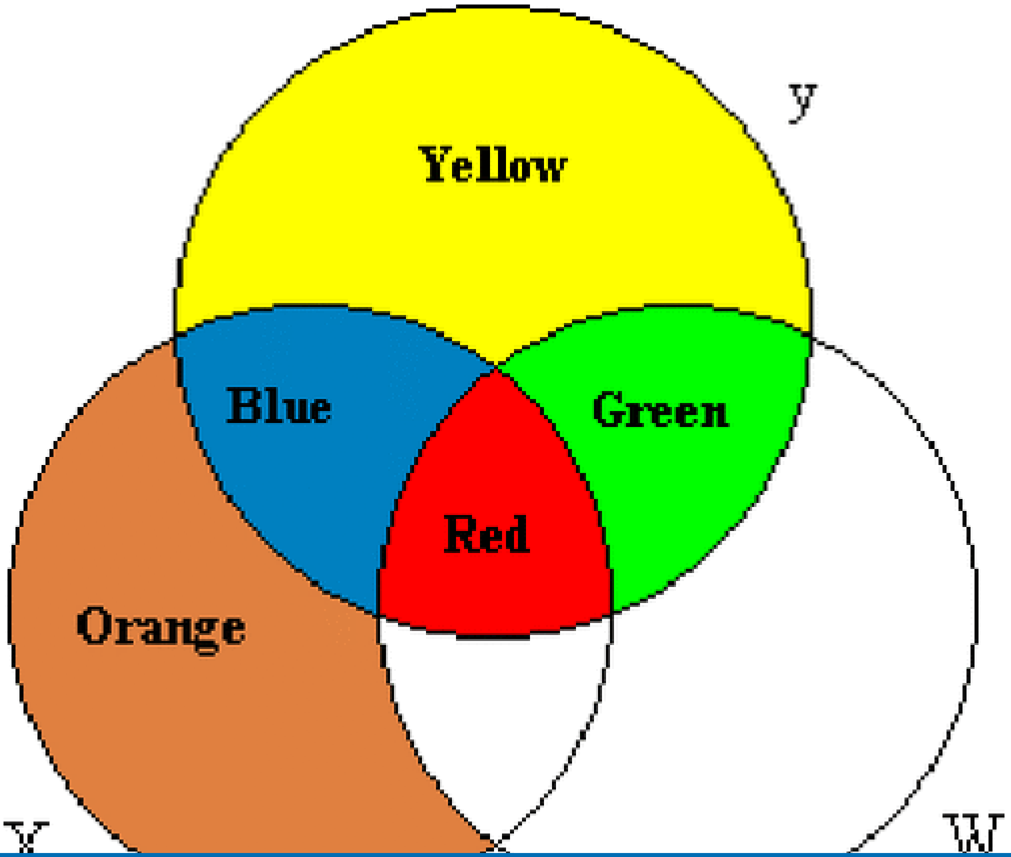
a. The purple area represents information used by the OLS formula when estimating β_X; if this information corresponds to variation in y uniquely explained by variation in X, the resulting estimate of β_X is unbiased.

b. A larger purple area means that more information is used in estimation, implying a smaller variance of the β_X estimate.

c. The black area, the variation in y that cannot be explained by X, is attributed by OLS to the error term, and so the magnitude of this area represents the magnitude of the OLS estimate of σ^2, the variance of the error term. Notice a subtlety here. The purple area is interpreted as information; the black area interpretation is quite different - its magnitude reflects the magnitude of a parameter estimate.

To analyze multiple regression, Kennedy adopted the three intersecting circles diagram of Cohen and Cohen (1975), which they named the “Ballantine” because of its resemblance to the logo of a brand of beer; Kennedy reinterpreted the areas as described above, and to emphasize this new interpretation changed its spelling to “Ballentine.” Such a diagram is shown in Figure 2, in which a new circle marked W is added (with an associated slope β_w), representing variation in another explanatory variable. In the presentations of Cohen and Cohen (1975) and Ip (2001), the overlap between the y circle and the X and W circles represents the variation in y explained by variation in X and in W. The ratio of this area (the blue plus red plus green area in Figure 2) to the y circle is interpreted as the R^2 from regressing y on X and W. Trouble happens in the presence of suppressor variables because the red area may have to be negative.

Figure 2. Ballentine Venn diagram.



Article contents
Related research

Display full size

Kennedy does not attempt to exposit R^2, but rather looks at a completely different set of OLS properties - bias and variance. The following section describes what in my experience are some particularly effective ways of using this diagram to teach the properties of OLS estimates in the CLR model.

3. Teaching some Properties of OLS

I begin by carefully explaining the interpretations of the areas in Figure 1 as described above. Following this I put up Figure 2 on the overhead and ask the class what will happen using OLS when there is more than one explanatory variable, drawing to their attention that it is not obvious what role is played by the red area. I note that if we were to regress y on X alone, OLS would use the information in the blue plus red areas to create its estimate of β_x, and if we were to regress y on W alone OLS would use the information in the green plus red areas to create its estimate of β_w. I present three options for the OLS estimator when y is regressed on X and W together.

a. Continue to use blue plus red to estimate β_x and green plus red to estimate β_w.

b. Throw away the red area and just use blue to estimate β_x and green to estimate β_w

c. Divide the red information into two parts in some way, and use blue plus part of red to estimate β_x and green plus the other part of red to estimate β_w.

I point out that several special cases of option c are possible, such as using blue and all of red to estimate β_x and only green to estimate β_w, or dividing red "equally" in some way.

After setting this up I inform students that they are to guess what OLS does, and ask them to vote for one of these options. (Voting has to be done one by one, because if the class at large is asked to vote, invariably nobody votes for anything; Kennedy (1978) is an exposition of this pedagogical device.) I have never had a majority vote for the correct answer. Next I ask the class why it would make sense for an estimating procedure to throw away the information in the red area. (It is this throwing away of the by the presence of a suppressor variable.) The ensuing discussion is quite useful, with good students eventually figuring out the following.

a. The information in the red area is bad information - it has y “moving together” with X and also with W so that we don't know if the movements in y are due to X or due to W, so to be on the safe side we should throw away this information.

b. If only the blue area information is used to estimate β_x and the green area information is used to estimate β_w, unbiased estimates are produced, because the blue area corresponds to variation in y uniquely attributable to X and the green area corresponds to variation in y uniquely attributable to W.

An instructor may wish to elaborate on point b by noting that the variation in y in the red area is actually due to joint movements in X and W because the red area corresponds to X and W “moving together” as well as to y and X moving together and y and W moving together. Suppose that in the red area when X changes by one W changes by two, so that a joint movement of one by X and two by W gives rise to a movement in y of $\beta_x + 2\beta_w$. If $\beta_x = 5$ and $\beta_w= 7$, this would be a movement of 19. If we were to match this 19 movement in y with a unit movement in X we would get a β_x estimate badly off the true value of $\beta_x = 5$. When this is combined with the unbiased estimate coming from the blue area information, a biased estimate results. Similarly, if the 19 movement in y were matched with a two movement in W we would get an estimate of 9.5 for β_w, badly off its true value of 7.

Instructors presenting an algebraic version of this material can demonstrate this result by working through the usual derivation of the OLS estimate of β_x as $(X^{*t}X^*)^{-1}X^*y^*$ where $y^* = M_wy$ and $X^* = M_wX$ with $M_w = I - W(W^tW)^{-1}W^t$. The residualizing matrix M_w removes that part of a variable explained by W, so that, in Figure 2, y^* and X^* are represented by areas blue plus yellow and orange plus blue, respectively; the OLS estimate results from using the information in their overlap, the blue area. This matrix formulation reveals how the case of three rather than two explanatory variables would be analyzed. Let X represent a single explanatory variable and W represent a matrix of observations on Z and Q, the other two explanatory variables. The W circle in the Venn diagram now represents the union of the Z and Q circles.

The instructor can finish by noting that the yellow area in Figure 2 represents the magnitude of σ^2, the variance of the error term. The OLS estimating procedure uses the

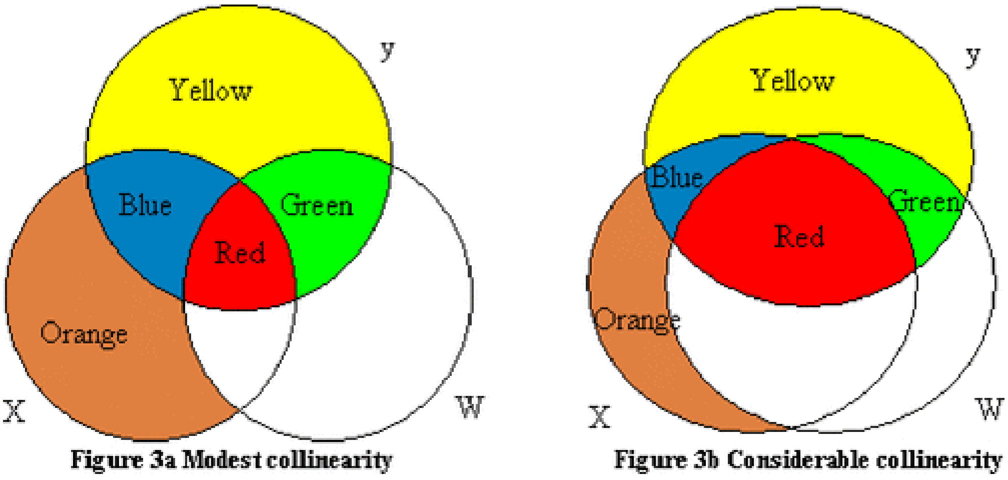
Article contents Related research

In this case this is the yellow area, which is the correct area, so the OLS estimator of σ^2 is unbiased. This sets up the Venn diagram for investigation of several features of OLS estimation, of which three are discussed below.

3.1 Multicollinearity

Ask the students how a greater degree of multicollinearity would manifest itself on the Venn diagram. They should be able to guess that it is captured by increasing the overlap between the X and W circles, as shown by moving from Figure 3a to Figure 3b. Next ask them if the higher collinearity causes bias. I go around the class and ask everyone to commit to yes, no, or don't know. (Those answering don't know are not asked later to explain their rationale.) Once this has been done, have someone who voted with the majority offer an explanation, and then have someone who voted with the minority offer a counter-explanation. Work with this until everyone sees that because the OLS formula continues to use the information in the blue and green areas to estimate the slopes of X and W, these estimates remain unbiased - the blue and green areas continue to correspond to variation in y uniquely attributable to X and W, respectively.

Figures 3a and 3b. Ballentine Venn diagrams displaying modest and considerable collinearity.


y
Yellow
Blue
Green
Red
Orange
X
W
Figure 3a Modest collinearity
y
Yellow
Blue
Green
Red
Orange
X
W
Figure 3b Considerable collinearity

Display full size

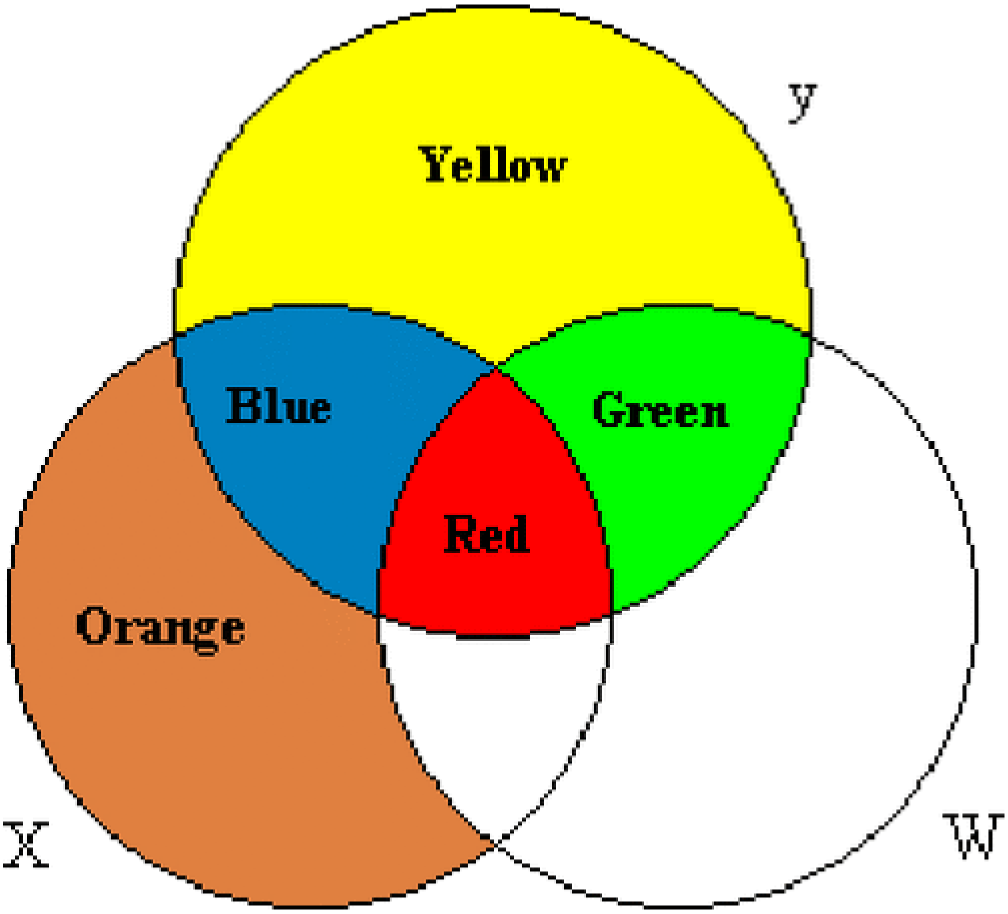
Next ask the students what the higher collinearity does to the variance of the estimate of β_x (or β_w). I go around the class and ask everyone to commit to increase, decrease, no change, or don't know. Direct a discussion as above until everyone sees that because the blue area shrinks in size, less information is used to estimate β_x, and so the variance of the OLS estimator of β_x is larger.

In summary, it should now be clear that higher collinearity increases variance, but does not cause bias. Finish by asking what happens in the diagram if X and W become perfectly collinear - the blue and green areas disappear and estimation is impossible.

3.2 Omitting a Relevant Explanatory Variable

Return to Figure 2, reproduced here as Figure 4, specifying that both X and W affect y, and ask how the properties of the OLS estimate of β_x would be affected if W is omitted from the regression, perhaps because a researcher isn't aware that it belongs, or has no way of measuring it. There are three questions of interest here.

Figure 4. Ballentine Venn diagram.



Display full size

First, is bias created whenever a relevant explanatory variable is omitted? I ask everyone to commit to yes, no, or don't know. After this voting, discussion should area to estimate β_X, and so is clearly biased because the red area is contaminated information. The direction of the bias cannot be determined from the diagram. Before leaving this the instructor can ask under what special circumstance would no bias be created? The students should be able to guess that when the X and W circles do not overlap (that is, they are orthogonal, such as would be the case in a designed experiment), there is no red area, so omission of a relevant explanatory variable will not create bias.

Second, what can we say about the variance of the OLS estimator compared to when W is included? I ask everyone to commit to larger, smaller, no change, or don't know. After this voting, the discussion should continue until everyone sees that because the blue plus red area information is used (instead of just the blue area information), more information is being used, so that variance should be smaller. What if the omitted explanatory variable W is orthogonal to X? In this case the OLS estimator continues to use just the blue information, so variance is unaffected.

At this stage the instructor might want to note that by omitting a relevant explanatory variable it should be clear that bias is created, a bad thing, but variance is reduced, a good thing, and comment that the mean square error criterion becomes of interest here because it is a way of trading off bias against variance. A good example to use here is the common procedure of dropping an explanatory variable if it is highly collinear with other explanatory variables. Ask the students how the results developed above could be used to defend this action. They should be able to deduce that omitting a highly-collinear variable can markedly reduce variance, and so may (but may not!) reduce the mean square error.

Third, what can we say about our estimate of σ^2 (the variance of the error term)? I ask everyone to commit to unbiased, biased upward, biased downward, or don't know. After this voting, the discussion should continue until everyone sees that the OLS procedure uses the magnitude of the yellow plus green area to estimate the magnitude of the yellow area, so the estimate will be biased upward. The instructor can follow up by asking if this bias disappears if the omitted explanatory variable W is orthogonal to X.

In summary, omission of a relevant explanatory variable in general biases coefficient estimates, reduces their variances, and causes an overestimate of the variance of the error term. If the omitted variable is orthogonal to the included variable, estimation

Article contents
Related research

3.3 Detrending Data

Suppose W is a time trend. How will the β_x estimate be affected if the time trend is removed from y and removed from X, and then detrended y is regressed on detrended X? This issue could be addressed by a voting procedure as described above, but I find it better to structure a computer assignment to illustrate this problem. Have the students obtain some data for y, X and W, ensuring that X and W are not orthogonal and that the OLS slope coefficients on X and W are substantive. I have them use 30 years of quarterly observations on a Canadian logged real monetary aggregate (y), logged real gross domestic product (X), and a logged interest rate (W). All data can be found in the CANSIM (www.statcan.ca/english/CANSIM) data base; data values have been logged to reflect the functional form specification usually adopted in this context.

They are then told to perform the following three estimations, following which they are asked to use the Ballentine Venn diagram to explain their results:

1. Regress y on X and W to obtain β_x, and its estimated variance vb^*.
2. Regress X on W, save the residuals r, and regress y on r to get c^*, the estimate of the r coefficient, and its estimated variance vc^*.
3. Regress y on W, save the residuals s, and regress s on r to get d^*, the estimate of the r coefficient, and its estimated variance vd^*.

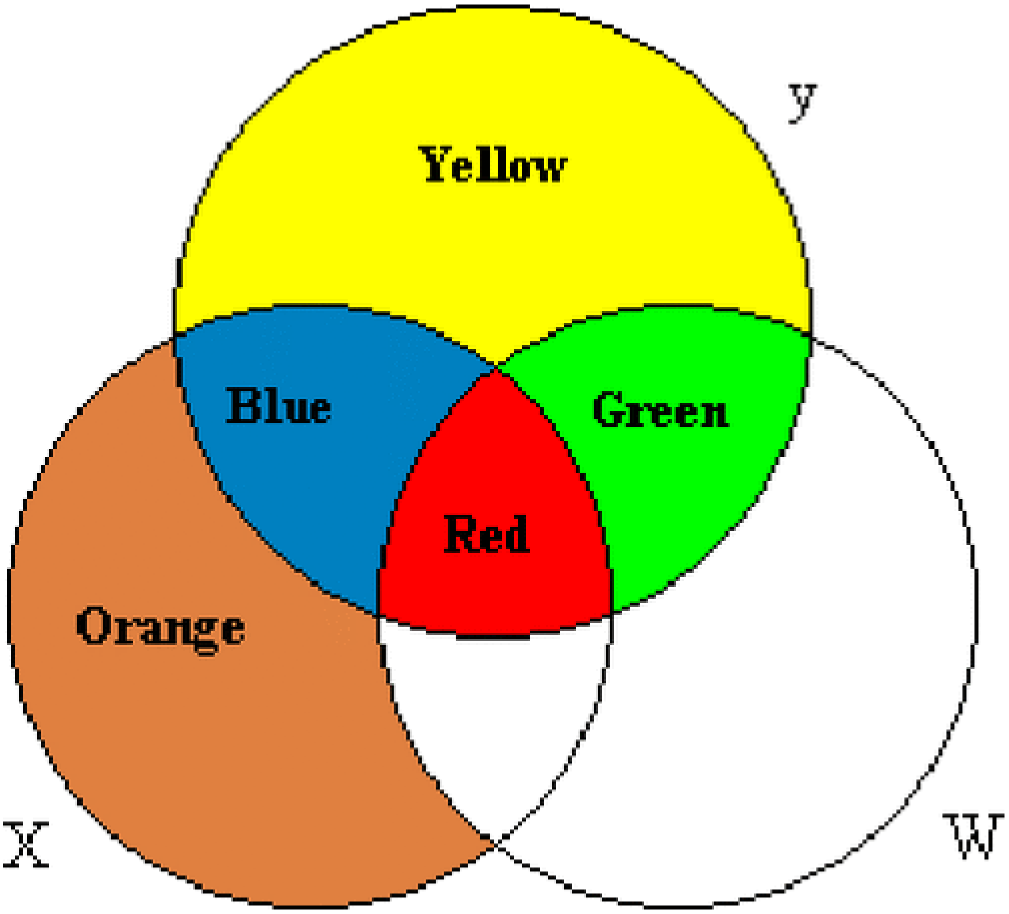
With their data my students obtain results reported in Table 1.

Table 1. Results from estimating with residualized data.

Download CSV Display Table

Students are surprised that these three estimates b^*, c^*, and d^* are identical to six decimal points. Most can employ the Ballentine to create an explanation for this. First, b^* is the usual OLS estimate, resulting from using the information in the blue area in Figure 5. Second, r is the part of X that cannot be explained by W, namely the orange plus blue area. The overlap of these two areas is the blue area, so regressing the y circle on the orange plus blue area uses the blue area information - exactly the same the part of y that cannot be explained by W, namely the blue plus yellow area. The overlap between s and r is the blue area, so regressing s on r (the blue plus yellow on the orange plus blue) uses the blue area information, once again exactly the same information as for estimating b^* and c^*. So this estimate should be identical to the other two.

Figure 5. Ballentine Venn diagram.



Display full size

Trouble begins when they try to explain the estimated variance results. Because exactly the same information is being used to produce b^*, c^* and d^*, they should all have exactly the same variance. But in Table 1 the three numbers are different. Students react to this in one of four different ways.

1. They ignore this problem, pretending that all they have to do is explain why the slope estimates are identical. Or they don't realize that the three variances are equal, and so believe that these differing numbers do not need comment.

2. They claim that all three numbers are identical except for rounding error

Article contents
Related research

3. They note that vb^* and vd^* are close enough that we can legitimately claim they are identical except for rounding error, note that vc^* is markedly higher, but are unable to offer an explanation.

4. They realize that although the true variances are equal, the estimated variances may not be equal, and use the Venn diagram to explain how this happens. When estimating b^* the variation in y not explained is the yellow area, so σ^2, the variance of the error term, is estimated by the magnitude of the yellow area. Similarly, when estimating d^* (by regressing blue plus yellow on orange plus blue) the variation in the dependent variable not explained is the yellow area, so in this case σ^2 is also estimated by the magnitude of the yellow area. But in the case of estimating c^* (by regressing the entire y circle on orange plus blue) the variation in y not explained is the yellow plus red plus green areas. As a result, in this case σ^2 is overestimated. This overestimation causes vc^* to be larger than vb^* and vd^*.

Instructors may wish to supplement this explanation by noting that the formula for the variance of an OLS estimator involves both σ^2 and variation in the explanatory variable data which is “independent” of variation in other explanatory variables. (In this example variance would be given by the formula $\sigma^2(X^tM_wX)^{-1}$.) In all three cases here, the “independent” variation in X is reflected by the blue plus orange areas, so the relative magnitudes of the estimated variances depend entirely on the estimates of σ^2.

How does all this relate to regressing on detrended data? If W is a time trend, then s is detrended y and r is detrended X, so that regressing s on r produces estimates identical to those of regressing on raw data. One concludes that it doesn't matter if one regresses on raw data including a time trend, or if one removes the linear trend from data and regresses on detrended data. Similarly, if W is a set of quarterly dummies, it doesn't matter if one regresses on raw data plus these dummies, or if one regresses on data that have been linearly deseasonalized. More generally, this reflects the well-known result that slope estimates are identical using raw data or appropriately residualized data.

4. Conclusion

The Ballentine Venn diagram is not new to the literature. Kennedy (1998) exposits the

☰ Article contents
Related research

irrelevant explanatory variable, and the rationale behind instrumental variable estimation. In Kennedy (1989), the Ballentine is used to exposit tests for non-nested hypotheses. The main contribution of this paper, beyond drawing readers' attention to this use of Venn diagrams, is to describe some particularly effective ways of using this diagram when teaching regression analysis.

There do exist special situations in which the Ballentine can mislead. For example, the Ballentine erroneously suggests that the OLS slope estimates from regressing y on X and W can be reproducing by the following procedure: obtain r by taking the residuals from regressing X on W, obtain v by taking the residuals from regressing W on X, and then regress y on r and v. My experience with its use in the classroom has been overwhelmingly positive, however; confined to standard analyses, the advantages of this Venn diagram interpretation as a pedagogical device are too powerful to ignore.

References

1. Cohen, J., and Cohen, P. (1975), Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences, Hillside, NJ: Lawrence Erlbaum Associates.

   Google Scholar

2. Ip, E. H. S. (2001), “Visualizing Multiple Regression,” Journal of Statistics Education [Online], 9 (1). (ww2.amstat.org/publications/jse/v9n1/ip.html)

   Google Scholar

3. Kennedy, P. E. (1978), “Democlass: A Variation on the Question/Answer Technique,” Journal of Economic Education, 9, 128–130.

   Web of Science ® | Google Scholar

4. Kennedy, P. E. (1981), “The ‘Ballentine’: A Graphical Aid for Econometrics,” Australian Economic Papers, 20, 414–416.

   Web of Science ® | Google Scholar

☰ Article contents | Related research

5. Kennedy, P. E. (1989), “A Graphical Exposition of Tests for Non-nested Hypotheses,” Australian Economic Papers, 28, 160–165.

Google Scholar

6. Kennedy, P. E. (1998), A Guide to Econometrics (4th ed.), Cambridge, MA: MIT Press.

Google Scholar

Download PDF

Related research

People also read

Recommended articles

Cited by
1

Information for

Authors

R&D professionals

Editors

Librarians

Societies

Open access

Overview

Open journals

Open Select

Dove Medical Press

F1000Research

Opportunities

Reprints and e-prints

Advertising solutions

Accelerated publication

Corporate access solutions

Help and information

Help and contact

Newsroom

All journals

Books

Keep up to date

Register to receive personalised research and resources by email

Sign me up

Privacy policy Cookies Terms & conditions

Accessibility

Taylor and Francis Group

Registered in England & Wales No. 01072954
5 Howick Place | London | SW1P 1WG

Article contents Related research