Journal of Statistics Education ›
Volume 10, 2002 - Issue 1

Original Articles

More on Venn Diagrams for Regression

E. Kennedy Peter

| Published online: 01 Dec 2017

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

Abstract

A Venn d… of coefficient estimate… ns for how it can be u… raying results a… ue is not compror…

Detrending…

1. Intr…

In a rece… ms for enhanci… ote is to

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.

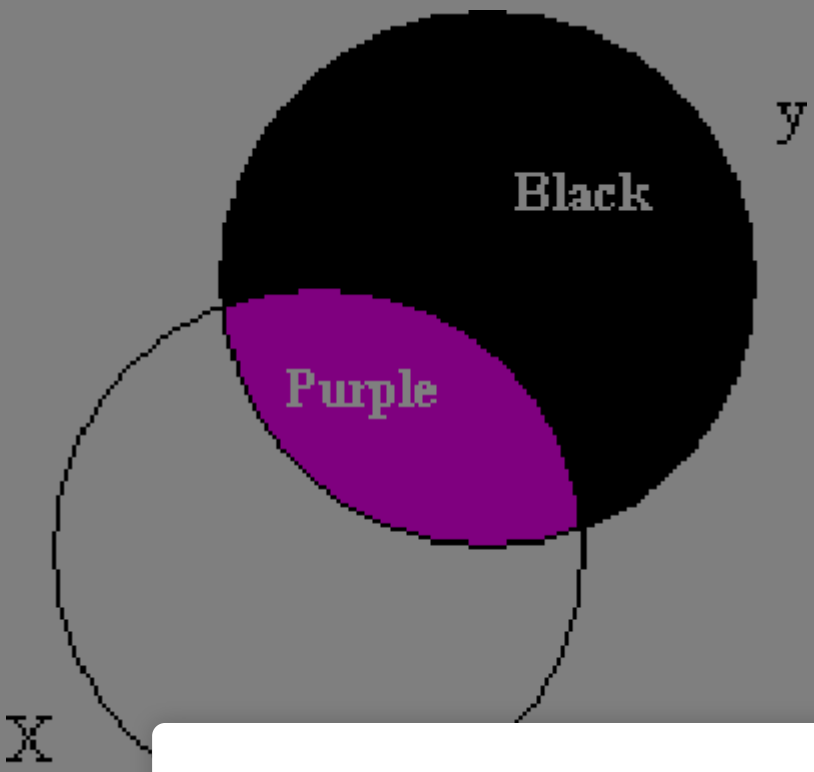
Ip's article is confined to the use of Venn diagrams for analyzing the coefficient of determination R^2, partial correlation, and sums of squares. In these cases, exposition is compromised in the presence of suppressor variables. But there are other concepts in regression analysis, thought by many to be of considerably more importance than R^2, which ar ng bias and variance tion of Venn diagram these diagram

2. An A

Kennedy ariance in the cont + ε. Here the depe es represen d error ε. The usua d that X is fixed in e measured as devia diagrams

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 ful

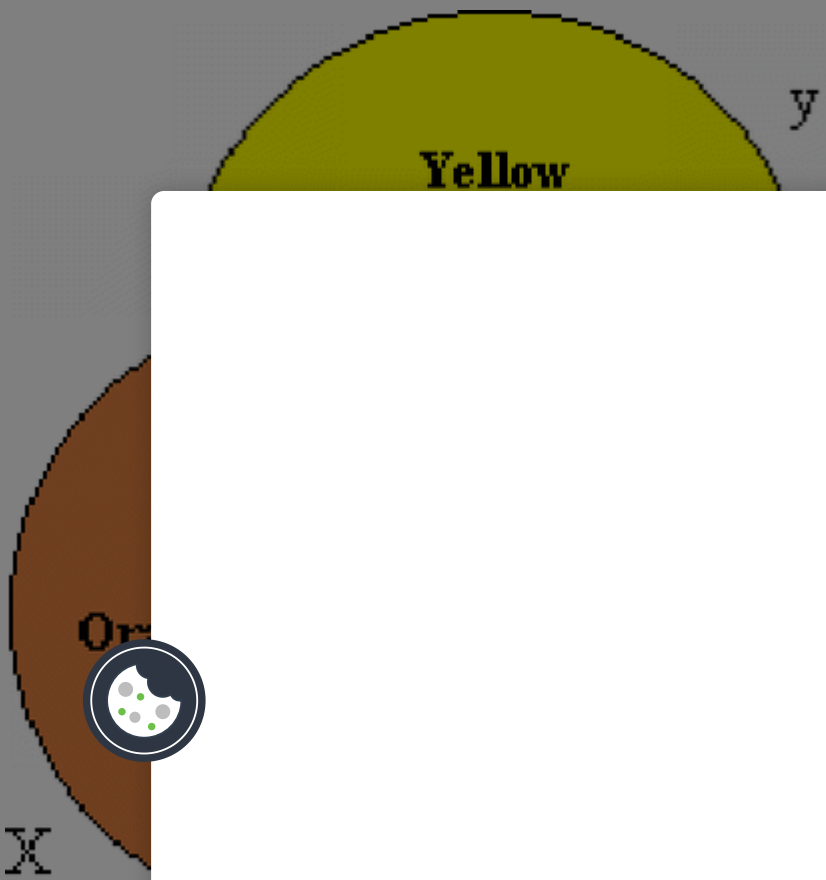
Although as in Ip's applicati

- The pu timating β_x variation in X),
- A large mplying a smalle
- The bl ted by OLS to nagnitude

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.



Display ful

Kennedy ... fferent set

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.

- Continue to use blue plus red to estimate β_x and green plus red to estimate β_w.
- Throw away the red area and just use blue to estimate β_x and green to estimate β_w
- 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 o blue and all of red to ” in some way.

After set , and ask them to ecause if the class nnedy (197 rity vote for the c mating procedu away of the red area mpromised by the p seful, with good stu

- The in ther” with

to W, so to be on the safe side we should throw away this information.

- 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.

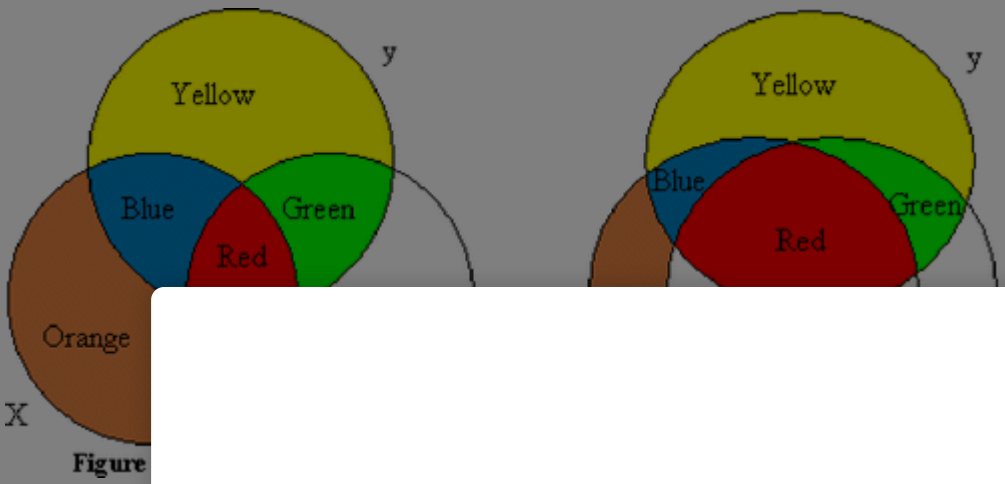
An instructor may wish to elaborate on point b by noting that the variation in y in the red area is actually due to joint movements in X and W because the red area corresponds to X and W "moving together" as well as to y and X moving together and y and W moving together. Suppose that in the red area 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 matrix M_w removes X^* are represer ne OLS estimate This matrix formulat ables would be analy a matrix of observat in the Venn diagr

The inst ts the magnitu ure uses the magnitu imate of σ^2. In this ca imator of σ^2 is unbias ures of OLS estimati

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.



Figure

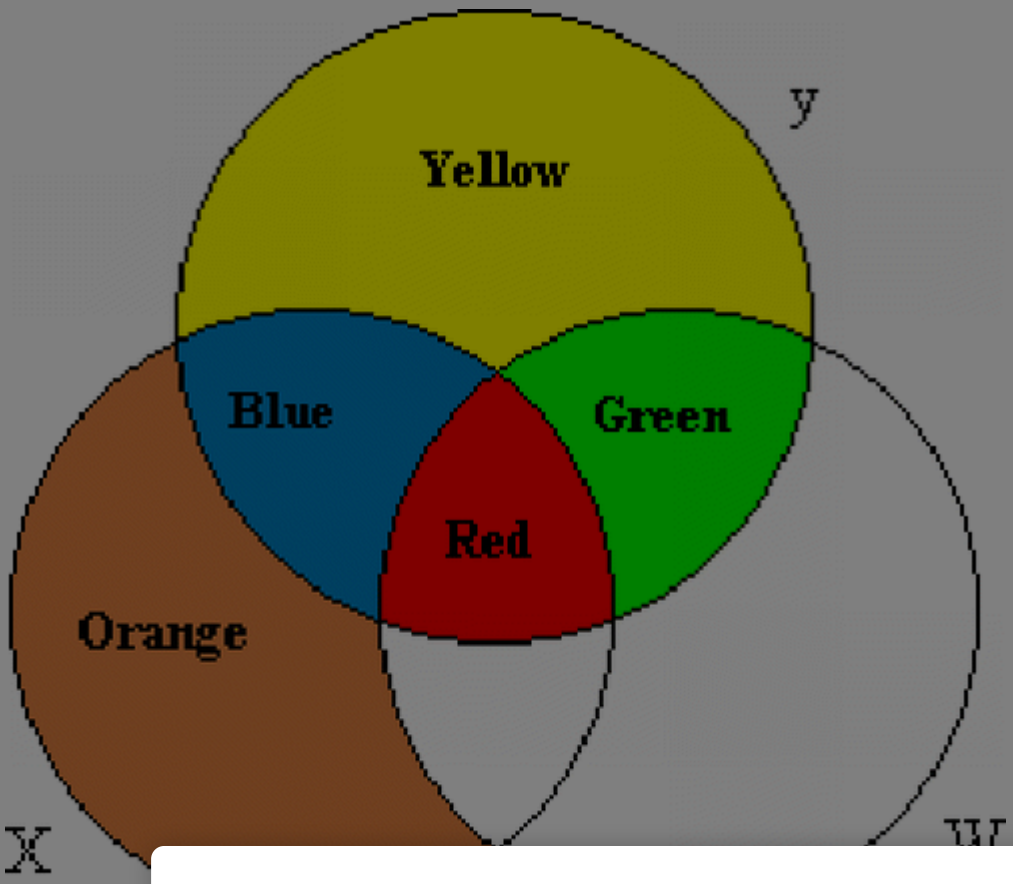
Display ful

Next ask ne estimate of β_x (or decrease, no ch that becau $_x$, and so the varia

In summ ce, but does not caus come perfectly npossible.

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.



First, is b ask
everyon should
continue blue plus red
area to e minated
infor m. Before
leavin no bias be
created? es do not
overlap ed
experim iable will
not crea

Second, d to when W

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

gonal to X.

In summ

oefficient

estimate

ance of the

error ter

timation

remains

mated.

3.3 Det

Suppo

trend is

removed

detrended

X? This i

, but I find it

better to

he students

obtain s

and that the

OLS slop

s of

quarterly

logged real

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^*.

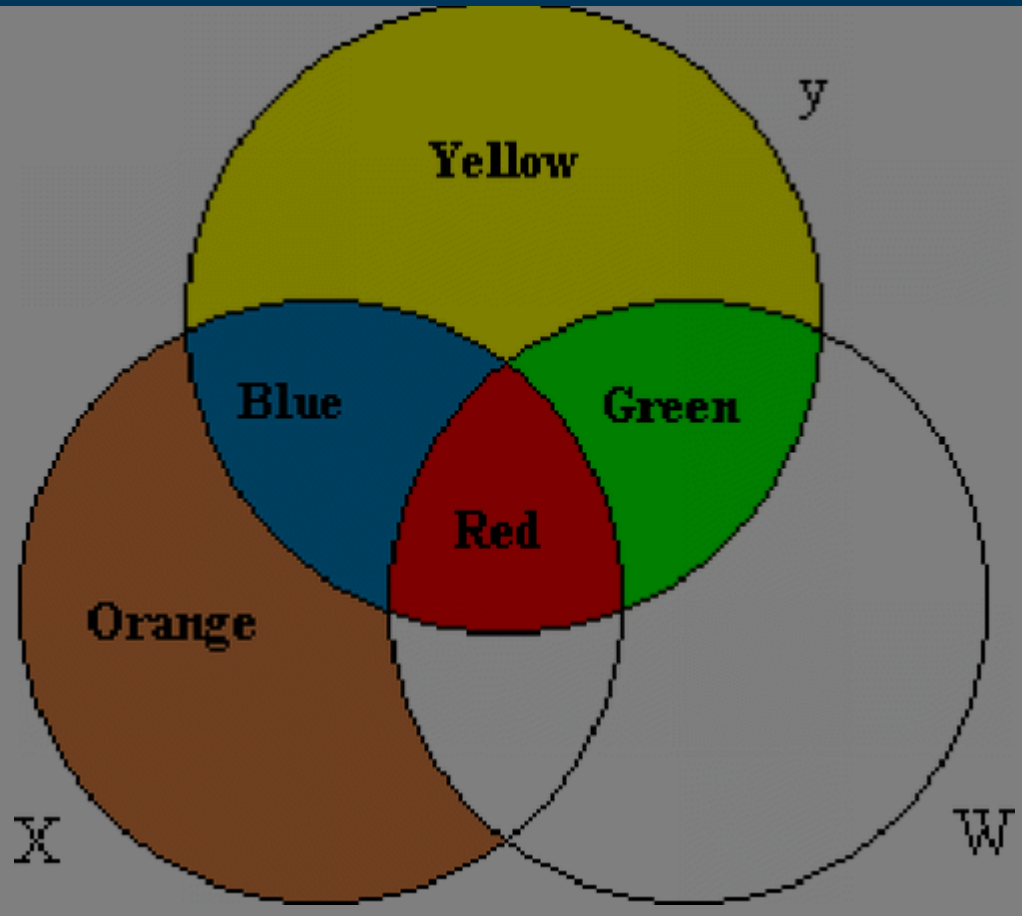
With their data my students obtain results reported in Table 1.

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 this. First, b^* is the e area in Figure 5. he orange plus blue g the y circle on the same informat d third, s is the part ea. The overl yellow on the c e same informat to the other two.

Figure 5.



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

1. They i hy the slope es are equal,
2. They c
3. The m they are ut are unable
4. They r riances may n . When estima e variance of the arly, when

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 residuali

4. Con

The Ball xposits the applicati ing an irrele ble estima n-nested hypothe ttention to this use f using this diagram

There d ample, the Ballentir ng y on X

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,” Journa

4. Kenne ,” Australian Econo

5. Ke otheses,” Austra

6. Kenne MIT Press.

   Googl

Download PDF

Related research

People also read | Recommended articles | Cited by 1

Information for

- Authors
- R&D professionals
- Editors
- Librarians
- Societies

Opportunities

- Reprints and e-prints
- Advertising solutions
- Accelerated publication
- Corporate access solutions

Open access

- Overview
- Open journals
- Open Select
- Dove Medical Press
- F1000Research

Help and information

- Help and contact
- Newsroom
- All journals
- Books

Keep up

Register t
by email

Sign

Copyright

Accessibi

Registered
5 Howick Pl

or & Francis Group
orma business

Article contents | Related research