The Stata Journal (*yyyy*) *vv*, Number *ii*, pp. 1–18

VizTest: Optimizing Confidence Intervals for Visual Testing

David A. Armstrong II
Western University
Department of Political Science
London, ON, Canada
dave.armstrong@uwo.ca

William Poirier
Western University
Department of Political Science
London, ON, Canada
wpoirier@uwo.ca

Abstract. For nearly a century, researchers have been aware of the problems inherent in visual testing—making inferences about the statistical differences in estimates based on the (non-)overlaps in their confidence intervals. A long line of research has identified how to ensure that a pair of intervals overlaps under the null hypothesis with a desired probability based on the ratio of the standard errors of the estimates and their correlation (Browne 1979; Afshartous and Preston 2010; Radean 2023). The solution works for a single pair of intervals, but except in the narrowest set of real-world circumstances, this solution applied to multiple pairs of intervals will not produce a single, predetermined type I error rate for all tests. Armstrong II and Poirier (Forthcoming) propose a novel solution to this problem that separates the pairwise tests of difference from the visualization. The `VizTest` package discussed here implements their suggestions.

Keywords: st0001, Confidence Intervals, Hypothesis Testing, Pairwise Comparisons, `viztest`

1 Introduction

Over the past century, confidence intervals have become ubiquitous in the presentation of statistical results. In fact, shortly after what appears to be among the first presentations of means and $\pm 2SE$ bars by Dice and Laraas (1936), Simpson and Roe (1939, p. 316) said of the method that it “promises to be one of the most important of numerical methods in zoology.” Confidence intervals have become one of the most used methods for presenting statistical results by scientists of all stripes.

Despite their utility and prevalence, confidence intervals still fall short of their potential. When confronted with a set of estimates and confidence intervals, viewers are often drawn to try to make comparisons about the similarities and differences of the underlying estimates—what we refer to as *visual testing*. In some cases this operation works; when confidence intervals do not overlap, the underlying estimates are statistically different from zero. However, when intervals overlap, the converse is not necessarily true. Scholars have developed rules of thumb for making these comparisons, identifying important features of the underlying estimates that animate these comparisons (Dice and Laraas 1936; Simpson and Roe 1939; Browne 1979). Others tried to formalize the process of visual testing such that for a pair of estimates the (non-)overlap in their confidence intervals would be a reliable indicator of statistical significance (Tukey 1991; Payton et al. 2003; Afshartous and Preston 2010; Radean 2023). While these techniques work as intended on a single pair of intervals, their utility wanes as the number of estimates increases.

Armstrong II and Poirier (Forthcoming) propose a different solution that decouples the pairwise hypothesis tests from the visualization. That is, they propose a method that performs pairwise comparisons numerically and then finds a visual representation that corresponds as closely as possible with the results of the tests. We briefly summarize the problem and solution that Armstrong II and Poirier (Forthcoming) identify. We then describe Stata software that implements these suggestions and describe its use in several examples.

2 The Visual Testing Problem and Previous Solutions

Confidence intervals are used by scientists across disciplines to present parameter estimates or model predictions along with their measures of their sampling or posterior uncertainty. In the frequentist case, confidence intervals convey point estimates and sampling uncertainty in a way that is easily engaged by readers. Armstrong II and Poirier (Forthcoming) suggest that when confronted with a collection of point estimates and confidence intervals, readers may desire to do the following:

- Test each estimate independently against a point null hypothesis (often, but not necessarily, zero).
- Evaluate the relative size of sampling uncertainty for different estimates.
- Identify whether a pair of estimates is statistically distinguishable.

The first two tasks are easily accomplished with $(1-\alpha)\times 100\%$ confidence intervals.[1] For the first task, if the null hypothesized value is excluded from the $(1-\alpha)\times 100\%$ confidence interval, then the null hypothesis is rejected at level α in favor of the two-sided alternative. Otherwise, the null hypothesis is not rejected. The second task is cognitively more complicated, but not impossible. The relative size of the sampling uncertainty for any pair of intervals would require estimating the length of each interval and then dividing one length by the other. Again, this is a task that is less reliably executed than the first but is feasible. Generally speaking, $(1-\alpha)\times 100\%$ confidence intervals cannot be used to perform tests of difference at the α level. We discuss the reasons for this below.

First, we consider the cases where comparisons are possible. For confidence intervals that do not overlap, we know that their estimates are statistically different from each other. For the converse to be true, we would need it to be the case that the intervals do not overlap, but the pairwise test would indicate an insignificant difference. Assuming,

[1] Here, α is the desired type I error rate of the test.

for the moment, that $\bar{y}$ is the larger estimate, $\bar{x}$ is the smaller estimate, and $\mathbf{V}$ holds their sampling variances and covariance. For confidence intervals for $\bar{x}$ and $\bar{y}$ to not overlap *and* their difference be statistically insignificant, both of the following would have to be true:

- The lower bound of the larger estimate is larger than the upper bound of the smaller estimate:
$$\bar{x} + q_\alpha s_{\bar{x}} < \bar{y} - q_\alpha s_{\bar{y}} \tag{1}$$
- The difference is smaller than the critical value required for significance:
$$\frac{\bar{y} - \bar{x}}{\sqrt{s_{\bar{x}}^2 + s_{\bar{y}}^2 - 2v_{\bar{x},\bar{y}}}} < q_\alpha \tag{2}$$

where $x < y$, $0 < s_{\bar{x}}$, $0 < s_{\bar{y}}$, $v_{\bar{x},\bar{y}}$ is the covariance of $\bar{x}$ and $\bar{y}$, and $q = F^{-1}\left(1 - \frac{\alpha}{2}\right)$. This implies:

$$s_{\bar{x}}^2 + s_{\bar{y}}^2 + 2s_x s_y < s_{\bar{x}}^2 + s_{\bar{y}}^2 - 2v_{\bar{x},\bar{y}}$$

or, crucially that $s_x s_y < -v_{\bar{x},\bar{y}}$. Since $v_{\bar{x},\bar{y}}$ is a covariance, it is bound in the range $[-s_x s_y, s_x s_y]$ and thus both propositions cannot be true simultaneously. This simply formalizes what we know to be true—estimates with non-overlapping confidence intervals are statistically different from each other.

Unfortunately, the converse of the above statement is not true. $(1 - \alpha) \times 100\%$ confidence intervals can overlap yet the estimates may still be statistically different from each other. For estimates with overlapping intervals to be statistically different from each other, the following two things would need to happen simultaneously:

- The lower bound of the larger estimate is smaller than the upper bound of the smaller estimate:
$$\bar{x} + q_\alpha s_{\bar{x}} > \bar{y} - q_\alpha s_{\bar{y}} \tag{3}$$
- The difference is greater than the critical value required for significance:
$$\frac{\bar{y} - \bar{x}}{\sqrt{s_{\bar{x}}^2 + s_{\bar{y}}^2 - 2v_{\bar{x},\bar{y}}}} > q_\alpha \tag{4}$$

Both inequalities hold when $s_{\bar{x}} s_{\bar{y}} > -v_{\bar{x},\bar{y}}$. Again, this simply formalizes what we already know—overlapping confidence intervals do not necessarily indicate insignificant differences in estimates.

To mitigate this inferential problem, researchers have tried to identify what Tyron (2001) called an *inferential confidence level* (we could call the inferential type I error rate γ). The main idea is that $(1-\gamma)\times 100\%$ confidence intervals will overlap under the null hypothesis with probability $(1-\alpha)$, thus rendering their (non)-overlaps a reliable test of statistical difference. Afshartous and Preston (2010) show that the probability two confidence intervals will overlap under the null hypothesis is:

$$\Pr(\text{Overlap}) = 2\left(1 - F\left(Z_\gamma \frac{\theta}{\sqrt{\theta^2+1-2\rho\theta}} + \frac{\frac{1}{\theta}}{\sqrt{1+\theta^{-2}-2\rho\theta^{-1}}}\right)\right) \tag{5}$$

where, Z_γ, the multiplier applied to the standard error in the inferential confidence interval is defined as:

$$Z_\gamma = F^{-1}\left(1-\frac{\gamma}{2}\right) \tag{6}$$

$$= \left[\frac{F^{-1}\left(\frac{1-\alpha}{2}\right)}{\frac{\theta}{\sqrt{\theta^2+1-2\rho\theta}} + \frac{\frac{1}{\theta}}{\sqrt{1+\theta^{-2}-2\rho\theta^{-1}}}}\right] \tag{7}$$

Figure 1 shows the inferential confidence levels $(1-\gamma)$ required to produce a 95% overlap in confidence intervals under the null hypothesis for hypothetical values of θ (the ratio of the standard errors of the two estimates) and ρ (the correlation between the two estimates). Generally, the 95% interval only works when there is a very uneven ratio of standard errors or an extreme negative correlation among the parameters. This also makes it clear that Tukey's (1991) solution of 84% confidence intervals works in some cases, but is not a general solution as in many cases 84% intervals will give the wrong answer.

While we could do this reliably for one pair of intervals, with many intervals, it is possible, perhaps even likely, that the ratios of their standard errors and their correlations will vary resulting in different required inferential confidence intervals. Further, consider three estimates, $\bar{x}, \bar{y}, \bar{z}$. It could be that the optimal inferential confidence level for the comparison of $\bar{x}$ and $\bar{y}$ is, say, .8 and the optimal inferential confidence level for the comparison of $\bar{x}$ and $\bar{z}$ is .7. What value should we use when drawing the inferential confidence interval for $\bar{x}$? One method would be to average these different values (or the Z_γ values that give rise to them), but that will result in both tests being at least slightly wrong.

3 Inferential Confidence Intervals for Visual Testing

Armstrong II and Poirier (Forthcoming) suggest an entirely different approach. Rather than relying on the overlap in confidence intervals to do the test, they suggest a three-step procedure.

1. Calculate all the pairwise tests using z-tests, t-tests, or indeed whatever kind of

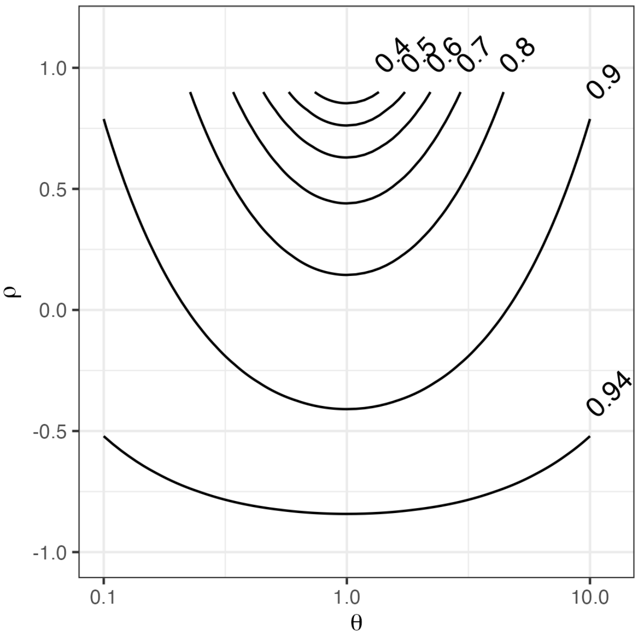

Figure 1: Confidence Levels Required for 95% Overlap

test is deemed appropriate. The tests are saved in the vector s_{ij} indicating whether the pairwise difference between estimates i and j is statistically significant (1) or not (0).

2. For a test level of γ, call it γ_0, calculate the $(1-\gamma_0)\times 100\%$ inferential confidence intervals. For all pairs of intervals identify whether they overlap or not. Store the results in the vector s^*_{ij} indicating whether intervals for estimates i and j are disjoint (1) or overlap(0).

3. Do a grid search over many candidate levels of γ_0, each time calculating the correspondence of s_{ij} and s^*_{ij} for all $i < j$. Specifically, we want to choose γ as follows:

$$\arg\max_{\gamma} \sum_{j=2}^{J}\sum_{i=1}^{j-1} I(s_{ij} = s^*_{ij}) \tag{8}$$

where $I(\cdot)$ is an indicator function that evaluates to 1 if the expression inside is true and zero otherwise. The value(s) of γ that maximize this expression are all equally good in terms of capturing the pairwise test results with the (non-)overlaps in their inferential confidence intervals.

Since the tests are performed independently of the intervals, the probability that the intervals overlap under the null hypothesis is incidental. As long as the (non-)overlaps in the intervals correspond with the (in)significance of the statistical tests, then the display conveys all the relevant information required for those viewing the display to test the equality of parameters.

3.1 `viztest` and the Reference Category Problem

Armstrong II and Poirier (Forthcoming) discuss visual testing as it relates to displays of model coefficients, predicted probabilities, first differences/marginal effects, or descriptive quantities such as means or proportions. One of the problems with using confidence intervals to do visual testing is that information about the covariance of the parameters is neither available in nor accounted for by the display. The so-called reference category problem is a special case of this.

The reference category problem arises when a categorical variable is represented in a regression model with a series of dummy variables, one level must be omitted as the reference category to prevent perfect collinearity. The coefficients for the non-reference groups convey the difference in the linear prediction between the reference group and each of the non-reference groups. However, testing the difference between two non-reference groups requires information about the covariance of the two parameter estimates.[2]

The function `viztest` can also solve the problem of reference category, in that it can operate on the output of `margins`. Using `margins` to calculate the predicted values for a categorical covariate in the model and then subjecting those values to `viztest` will provide a set of confidence intervals that can be used to test across any pair of values in the categorical variable. We provide an example of this below.

4 The `viztest` Command

4.1 Syntax

The `viztest` command performs the three steps mentioned above and presents the results in a user-friendly manner. The command syntax is as follows:

```
viztest [, lev1(#) lev2(#) incr(#) a(#) adjust(string) inc0 remc
   usemargins saving(string) ]
```

There are no required arguments, by default `viztest()` will operate on the `e(b)` and `e(V)` to find the inferential confidence intervals that correspond most closely with

[2] Let's say we wanted to test $H_0 : \beta_2 - \beta_1 = 0$ relative to the two-sided alternative. We would need to construct a t-statistic: $t = \frac{b_2 - b_1}{SE(b_2 - b_1)}$. Expanding the denominator, we would get $t = \frac{b_2 - b_1}{\sqrt{\text{var}(b_1) + \text{var}(b_2) - 2\text{cov}(b_1, b_2)}}$.

the normal-theory pairwise tests.

4.2 Options

`lev1(#)` A real value in the range (0,1) giving the confidence level that marks the lower bound of the grid search for acceptable inferential confidence level(s). The default is .25.

`lev2(#)` A real value in the range (0,1) giving the confidence level that marks the upper bound of the grid search for acceptable inferential confidence level(s). The default is .99.

`incr(#)` A real value in the range (0,1) giving the step size of the grid search. The algorithm will search from `lev1` to `lev2` in increments of `incr` to find the optimal inferential confidence level(s). The default is 0.01.

`a(#)` The p-value used to mark significance of the pairwise tests. The default is 0.05.

`adjust(string)` A string giving the multiplicity adjustment to use in the calculation of the pairwise test p-values. Possible values are: "bonferroni", "holm", "hochberg", "hommel", "bh", "by", "none". The default is "none".

`inc0` If `inc0` is issued as an argument, inferential confidence intervals will also include all tests of the parameters relative to zero—the univariate null hypothesis test. By default, zero is not included in the test.

`remc` If `remc` is issued, the last estimate will be removed from the analysis. Generally, this is the model constant, but, for example, in `margins`, the last estimate will not be the constant, so this option should not be invoked in that situation.

`usemargins` If `usemargins` is issued, `viztest` will use `r(b)` and `r(V)` returned from `margins` as the estimates and variances rather than `e(b)` and `e(V)` returned from the model.

`saving(string)` A string giving the stem of a filename. If specified, two files will be created. ∗_results.dta will hold the results of the grid search over all the levels. ∗_miss.dta will hold information about the pairwise tests that were not appropriately captured by the inferential confidence intervals, if any.

4.3 Returns

The function returns some values that may be useful to use in other contexts. It returns the four levels printed in the output as scalars as well as the results of the grid search as a matrix and a matrix of missed tests, if any exist. If no tests were missed, no returned value will exist.

Scalars

`r(smallest)`	smallest confidence level	`r(biggest)`	largest confidence level
`r(middle)`	median confidence level	`r(easiest)`	easiest confidence level

Matrices

`r(grid)`	grid search results	`r(miss_tests)`	tests not captured by inferential confidence intervals (if any)

5 Examples

In the following examples, we show how `viztest` works with both model objects and `margins`. The function works by capturing the estimated parameters in `e(b)` and variance-covariance matrix in `e(V)` or `r(b)` and `r(V)` depending on the context. Since the returned values from models tend to be standardized (insofar as the parameter estimates are returned in `e(b)` and that variance-covariance matrix in `e(V)`) and `margins` operates on such a wide variety of models, the `viztest` function should work for most results generated in Stata. We describe a few prototypical cases here.

5.1 Descriptive Quantities

Perhaps most simply, `viztest` works for descriptive quantities using the `mean` command. Using the `census13` data from the web, we can calculate the average marriage rate by region along with 95% confidence intervals.

```
. webuse census13, clear
(1980 Census data by state)
. mean mr, over(region)
Mean estimation                                  Number of obs = 50

                 Mean   Std. err.     [95% conf. interval]

c.mrgrate@region
           NE     .012054   .0006035      .0108411    .0132668
      N Cntrl    .0139124   .0004687      .0129704    .0148544
        South    .0161761   .0009298      .0143077    .0180445
         West    .0307456   .0137568      .0031002     .058391
```

We run `viztest` on the result to identify the appropriate inferential confidence level(s). Any inferential confidence level between 81% and 88% will capture all the pairwise tests.

```
. viztest

Optimal Levels:

Smallest Level: .81
Middle Level: .84
Largest Level: .88
Easiest Level: .81
```

```
No missed tests!
```

We use the 81% level to make a plot of the means using `coefplot`.

```
. local lev = round(`r(easiest)´*100)

.
. coefplot, sort(., by(b)) level(`lev´) ///
>      xtitle("Average Marriage Rate" "`lev´% Inferential Confidence Intervals") ///
>          rename(c.mrgrate@1.region = "Northeast" c.mrgrate@2.region = "North Central" ///
>          c.mrgrate@3.region = "South" c.mrgrate@4.region = "West")
```

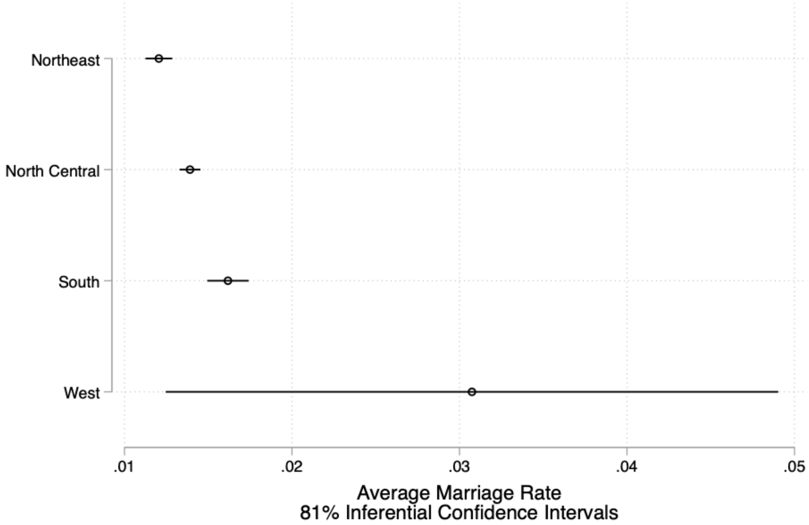

Figure 2: Plot of Means with 81% Inferential Confidence Intervals

A Note on Easiness

The output from `viztest` shows that all levels between 81% and 88% will appropriately capture the pairwise tests of interest. An important question is—which one of the levels is best. In Appendix 4 of their supplementary material, Armstrong II and Poirier (Forthcoming) have a discussion of the different options and what might be considered best. Values close to the lower end of the range will tend to accentuate significant differences (i.e., put more distance between the bounds of estimates that are different from each other) while making it more difficult to apprehend the insignificant differences (the confidence intervals of some insignificant differences will overlap only slightly). Choosing a value near the high end of the range will do the opposite—emphasize the insignificant differences at the expense of being able to identify insignificant differences.

As such, a good place to start might be the value in the middle of the upper and lower ends of the range.

Armstrong II and Poirier (Forthcoming) make a slightly different argument for cognitive easiness. They propose that there are two tests we should consider as being most important. The first is the pair of intervals for significantly different estimates whose upper and lower confidence bounds are closest to touching (i.e., the significant difference with the least space between their confidence intervals). The second is the pair of estimates that are not statistically distinguishable with the smallest overlap in their confidence intervals. The first test is more easily apprehended the smaller the confidence level we use. The converse is true for the second test—it is more easily apprehended as the confidence level increases. Armstrong II and Poirier (Forthcoming)'s easiness measure finds the level that makes these two tests equally easy to apprehend. If these tests are easy to see, then all others should be as well.

5.2 Regression Example

In the example below, we read in the `census13` data from the web and standardize all the independent variables we are going to use and change the labels so they are more amenable to making a coefficient plot.

```
. webuse census13, clear
(1980 Census data by state)
. egen z_mr = std(mrgrate)
. egen z_dr = std(dvcrate)
. egen z_ma = std(medage)
. egen z_pop = std(pop)
. label var z_mr "Marriage Rate"
. label var z_ma "Median Age"
. label var z_dr "Divorce Rate"
. label var z_pop "Population"
```

We run the regression of birth rate (`brate`) on the standardized variables that include a second degree polynomial in median age (`z_ma`).

```
. reg brate z_mr z_dr z_pop c.z_ma##c.z_ma
      Source |       SS           df       MS      Number of obs   =        50
-------------+----------------------------------   F(5, 44)        =     75.80
       Model |  37807.3612         5  7561.47224   Prob > F        =    0.0000
    Residual |  4389.45878        44  99.7604268   R-squared       =    0.8960
-------------+----------------------------------   Adj R-squared   =    0.8842
       Total |    42196.82        49  861.159592   Root MSE        =     9.988

------------------------------------------------------------------------------
       brate | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
        z_mr |  -3.477128   2.338042    -1.49   0.144    -8.189142    1.234887
        z_dr |   7.066246   2.399416     2.94   0.005      2.23054    11.90195
       z_pop |    .124971   1.533754     0.08   0.935    -2.966108     3.21605
        z_ma |  -23.84698   1.585808   -15.04   0.000    -27.04297   -20.65099
```

```
c.z_ma#c.z_ma |   4.315234   .7319742     5.90   0.000     2.840037    5.790431
              |
        _cons |   163.7111   1.584228   103.34   0.000     160.5183    166.9039
-------------------------------------------------------------------------------
```

There are two coefficients in the output where their confidence intervals overlap, yet they are statistically different from each other: `z_dr` and `z_pop`. A call to `test` shows that the difference is statistically significant.

```
. test z_dr = z_pop
 ( 1)  z_dr - z_pop = 0
       F(  1,    44) =    6.22
            Prob > F =    0.0165
```

We call `viztest` on the regression output invoking the `inc0` option to include all univariate tests against the null hypothesized value of zero. This shows that any inferential confidence level between 0.65 and 0.91 will produce confidence intervals whose (non-)overlaps correspond perfectly with the (in)significance of the pairwise differences. The inferential confidence level that makes the (lack of) differences easiest to apprehend is the 86% level.

```
. viztest, inc0

Optimal Levels:

Smallest Level: .86
Middle Level: .88
Largest Level: .91
Easiest Level: .86

No missed tests!
```

Replacing or Adding to Original Intervals

Armstrong II and Poirier (Forthcoming) suggest the use of inferential confidence intervals instead of existing $(1-\alpha) \times 100\%$ confidence intervals. They argue that three important tasks can still be executed with these new intervals—univariate tests against 0, pairwise tests of estimates, and the evaluation of relative sampling variability. One shortcoming of these (often smaller) intervals is that they do not provide the set of null hypothesis values that cannot be rejected at level α (as the $(1-\alpha) \times 100\%$ intervals do).

In addition, many potential users may be hesitant to replace their existing $(1-\alpha) \times 100\%$ confidence intervals with these often smaller ones for fear that peer reviewers will not be sympathetic to this approach. Authors who use smaller than generally accepted confidence intervals may be seen as trying to hide something. While we think people can use these tools in good faith with an appropriate discussion of their utility, we acknowledge that reviewers may continue to take a hostile position nonetheless. One easy solution is to add the inferential confidence intervals to a plot alongside the existing $(1-\alpha)\times 100\%$ intervals. This should ease concerns that reviewers have about hiding important aspects of the results while allowing users to make statistically reliable pairwise comparisons among estimates that would be otherwise challenging, if not impossible. We show an example of this in Figure 3 where we plot the inferential 86% intervals as thicker bars along with the original 95% intervals as thinner lines.

We use this new inferential confidence level to make a coefficient plot at the end of the code. You can see the output in Figure 3.

```
. local lev = round(`r(easiest)´*100)
. coefplot, drop(_cons) sort(., by(b)) level(95 `lev´) xline(0) ///
>   xtitle("Coefficient Estimates" "95% and `lev´% Confidence Intervals")
```

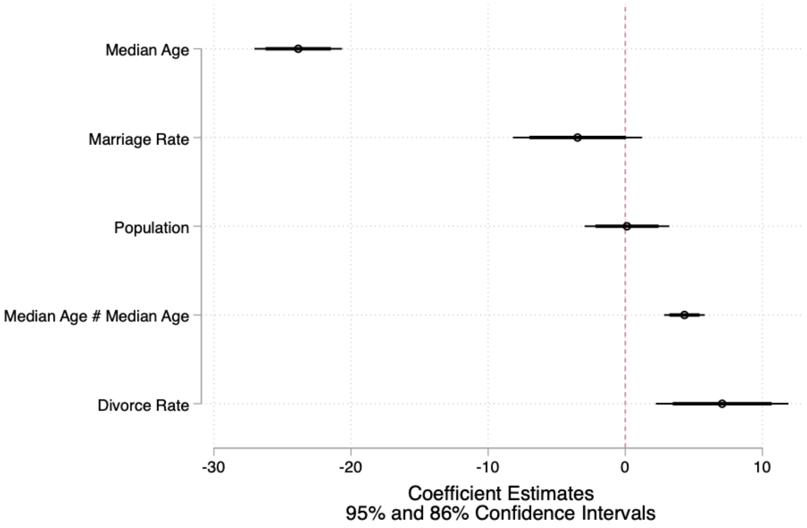

Figure 3: Coefficient Plot with 86% and the default 95% Confidence Intervals

5.3 Regression Example Using `margins`

In example 1, plotting the coefficients may not be particularly useful because the polynomial in median age makes it difficult to make the appropriate comparisons. Instead, we may want to plot the marginal effects (partial first differences) of the variables instead of their coefficients. To do this, we engage example 1 right after running the regression. We can then execute `margins` on the output.

```
. reg brate z_mr z_dr z_pop c.z_ma##c.z_ma
```

```
      Source |       SS           df       MS      Number of obs   =        50
-------------+----------------------------------   F(5, 44)        =     75.80
       Model |  37807.3612         5  7561.47224   Prob > F        =    0.0000
    Residual |  4389.45878        44  99.7604268   R-squared       =    0.8960
-------------+----------------------------------   Adj R-squared   =    0.8842
       Total |    42196.82        49  861.159592   Root MSE        =     9.988

-------------------------------------------------------------------------------
        brate | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
--------------+----------------------------------------------------------------
         z_mr |  -3.477128   2.338042    -1.49   0.144    -8.189142    1.234887
         z_dr |   7.066246   2.399416     2.94   0.005      2.23054    11.90195
        z_pop |    .124971   1.533754     0.08   0.935    -2.966108     3.21605
         z_ma |  -23.84698   1.585808   -15.04   0.000    -27.04297   -20.65099
              |
c.z_ma#c.z_ma |   4.315234   .7319742     5.90   0.000     2.840037    5.790431
              |
        _cons |   163.7111   1.584228   103.34   0.000     160.5183    166.9039
-------------------------------------------------------------------------------

. margins, dydx(*)

Average marginal effects                                    Number of obs = 50
Model VCE: OLS

Expression: Linear prediction, predict()
dy/dx wrt:  z_mr z_dr z_pop z_ma

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
        z_mr |  -3.477128   2.338042    -1.49   0.144    -8.189142    1.234887
        z_dr |   7.066246   2.399416     2.94   0.005      2.23054    11.90195
       z_pop |    .124971   1.533754     0.08   0.935    -2.966108     3.21605
        z_ma |  -23.84698   1.585808   -15.04   0.000    -27.04297   -20.65099
------------------------------------------------------------------------------
```

We execute `viztest` on the result, including zero in the tests and using the `usemargins` flag to ensure we use the returned results from the `margins` command.

```
. viztest, inc0 usemargins

Optimal Levels:

Smallest Level: .86
Middle Level: .88
Largest Level: .91
Easiest Level: .86

No missed tests!
```

The results here are the same, so we could use `marginsplot` to plot the marginal effect with the 86% inferential confidence intervals. The result is in Figure 4.

```
. local lev = round(`r(easiest)´*100)
.
. quietly margins, dydx(*)
. marginsplot, level(`lev´) recast(scatter) horizontal derivlabels ytitle("") ///
>   xtitle("Average Marginal Effect" "`lev´% Inferential Confidence Intervals") title("")
Variables that uniquely identify margins: _deriv
```

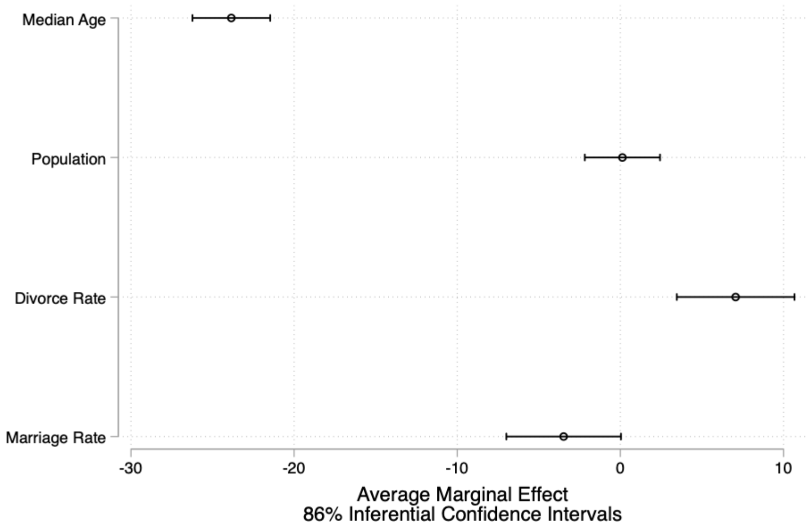

Figure 4: Margins Plot with 86% Inferential Confidence Intervals

5.4 Logistic Regression

Next, we turn to an example of logistic regression. There are interesting differences here relative to linear regression. The coefficients themselves are less interesting in these nonlinear models because the effects on the probability depend not only on the variable of interest, but on all variables in the model. Consider the `nhanes2d` data where we could predict high blood pressure `highbp` with a set of other variables including the size of place the patient lives `sizplace`. The place size variable is categorical with eight levels ranging from "Rural" to "Urbanized area; 3,000,000+" with the largest category being Rural.

```
. webuse nhanes2d, clear

. label def size2 1 "Urban: 3M+" 2 "Urban: 1M-3M" 3 "Urban: 250k-1M" ///
>        4 "Urban: <250k" 5 "Suburban: 25k +" 6 "Suburban: 10k-25k" ///
>        7 "Suburban: 2.5k-10k" 8 "Rural"

. label val sizplace size2

. logit highbp height weight age female i.sizplace

Iteration 0:  Log likelihood = -7050.7655
Iteration 1:  Log likelihood = -5831.4258
Iteration 2:  Log likelihood = -5816.4827
Iteration 3:  Log likelihood = -5816.4347
Iteration 4:  Log likelihood = -5816.4347

Logistic regression                                     Number of obs =  10,351
                                                        LR chi2(11)   = 2468.66
                                                        Prob > chi2   =  0.0000
Log likelihood = -5816.4347                             Pseudo R2     =  0.1751

---------------------------------------------------------------------------------------
             highbp | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
--------------------+------------------------------------------------------------------
             height |  -.0354691     .00367    -9.66   0.000    -.0426621   -.0282761
             weight |   .0502444   .0018393    27.32   0.000     .0466395    .0538492
                age |   .0472687   .0014668    32.23   0.000     .0443938    .0501435
             female |  -.3783304   .0644314    -5.87   0.000    -.5046136   -.2520472
                    |
           sizplace |
      Urban: 1M-3M  |  -.2623732   .0933404    -2.81   0.005    -.4453171   -.0794294
    Urban: 250k-1M  |  -.5584437   .0943781    -5.92   0.000    -.7434214   -.3734661
      Urban: <250k  |  -.1449434   .1007788    -1.44   0.150    -.3424662    .0525794
   Suburban: 25k +  |   -.144617   .1246263    -1.16   0.246      -.38888    .0996461
 Suburban: 10k-25k  |  -.1807813   .1241923    -1.46   0.145    -.4241937     .062631
Suburban: 2.5k-10k  |  -.1830208   .1044817    -1.75   0.080    -.3878012    .0217595
             Rural  |  -.3482749   .0779908    -4.47   0.000    -.5011341   -.1954157
                    |
              _cons |   .1525173   .6267083     0.24   0.808    -1.075808    1.380843
---------------------------------------------------------------------------------------

.
```

We see from the output above that the largest urbanized areas represent the reference group. High blood pressure is highest in the reference group, though the differences are not statistically significant for all comparisons. We see the reference category problem here as it is impossible for us to make many comparisons of interest among non-reference categories just using the model output. To make visible testing possible, we can first calculate the predictive margins of `sizplace`.

```
. margins sizplace

Predictive margins                                  Number of obs = 10,351
Model VCE: OIM

Expression: Pr(highbp), predict()

------------------------------------------------------------------------------
                   |            Delta-method
                   |     Margin   std. err.      z    P>|z|     [95% conf. interval]
-------------------+----------------------------------------------------------
          sizplace |
       Urban: 3M+  |   .4760551   .0132904    35.82   0.000     .4500064    .5021037
     Urban: 1M-3M  |   .4254659   .0121107    35.13   0.000     .4017294    .4492024
   Urban: 250k-1M  |   .3697427   .0119445    30.95   0.000     .3463318    .3931535
     Urban: <250k  |   .4480235   .0142368    31.47   0.000     .4201199    .4759271
   Suburban: 25k + |   .4480864   .0200776    22.32   0.000      .408735    .4874379
 Suburban: 10k-25k |   .4411205   .0199351    22.13   0.000     .4020484    .4801926
Suburban: 2.5k-10k |   .4406896   .0151407    29.11   0.000     .4110144    .4703649
            Rural  |      .4091   .0069815    58.60   0.000     .3954165    .4227835
------------------------------------------------------------------------------
```

The output from `viztest` shows that any level between 86% and 93% will represent all tests appropriately, but that the 90% intervals are the best. Again, we could present both the 95% and 90% intervals to permit visual testing, but also perhaps alleviate the discomfort of reviewers to presenting confidence bounds that are “too narrow”.

```
. viztest, usemargins

Optimal Levels:

Smallest Level: .86
Middle Level: .89
Largest Level: .93
Easiest Level: .9

No missed tests!
```

Figure 5 allows us to make valid inferences among any pair of values from the `sizplace` variable regardless of whether one of those values is for the reference group.

6 Conclusion

The `viztest` function is a simple function that allows users to identify confidence levels that permit reliable visible testing. These inferential confidence intervals can either replace default 95% intervals or be added on to an existing plot. The flexibility ensures that users can satisfy both reviewers who often balk at confidence intervals that are “too narrow” and readers who often would like to be able to compare estimates, but until now have lacked a viable way of doing so with existing model output and graphical presentations.

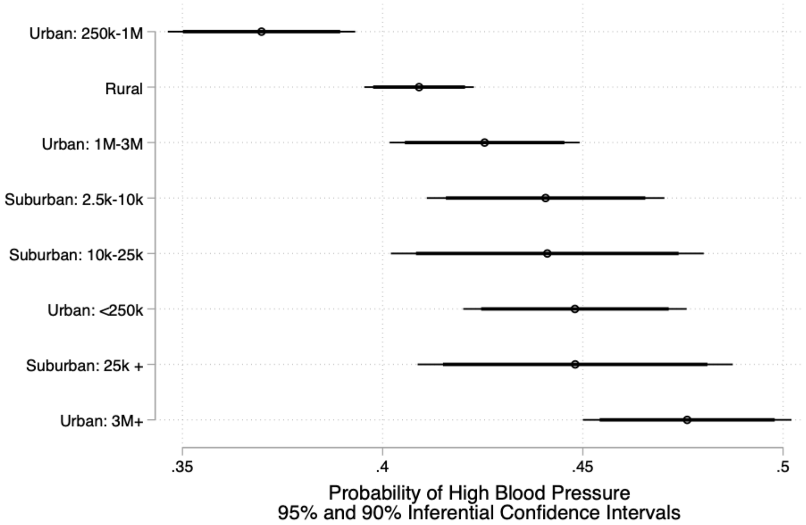

Figure 5: Effect Plot with 95% and 90% Inferential Confidence Intervals

About the authors

David A. Armstrong II is a Professor of Political Science and the Canada Research Chair in Political Methodology at Western University.

William Poirier is a Ph.D. student studying political methodology at Western University.

7 References

Afshartous, D., and R. A. Preston. 2010. Confidence Intervals for Dependent Data: Equating Non-overlap with Statistical Significance. *Computational Statistics and Data Analysis* 54: 2296–2305.

Armstrong II, D. A., and W. Poirier. Forthcoming. Decoupling Visualization and Testing when Presenting Confidence Intervals. *Political Analysis* .

Browne, R. H. 1979. On Visual Assessment of the Significance of a Mean Difference. *Biometrics* 35(3): 657–665.

Dice, L., and H. Laraas. 1936. A Graphic Method for Comparing Several Sets of Measurements. *Contributions from the Lab of Vertebrate Genetics* (3): 1–3.

Payton, M. E., M. H. Greenstone, and N. Schenker. 2003. Overlapping confidence intervals or standard error intervals: What do they mean in terms of statistical significance? *Journal of Insect Science* 3(1): 34.

Radean, M. 2023. The Significance of Differences Interval: Assessing the Statistical and Substantive Difference between Two Quantities of Interest. *Journal of Politics* 85(3): 969–983.

Simpson, G. G., and A. Roe. 1939. *Quantitative Zoology, revised edition*. New York, NY: McGraw-Hill.

Tukey, J. 1991. The Philosophy of Multiple Comparisons. *Statistical Science* 6(1): 100–116.

Tyron, W. W. 2001. Evaluating Statistical Difference, Equivalend and Indeterminacy using Inferential Confidence Intervals: An Integrated Alternative Method of Conducting Null Hypothesis Statistical Tests. *Psychological Methods* 6(4): 371–386.