factorplot: Improving Presentation of Simple Contrasts in GLMs

by David A. Armstrong II

Abstract Recent statistical literature has paid attention to the presentation of pairwise comparisons either from the point of view of the reference category problem in GLMs (e.g., Easton et. al., 1991; Firth and Menzes, 2004; Plummer, 2004) or in terms of multiple comparisons (e.g., Bretz et. al., 2010; Hsu, 1996). Both schools of thought are interested in the parsimonious presentation of sufficient information to enable readers to evaluate the significance of contrasts resulting from the inclusion of qualitative variables in GLMs. While considerable advances have been made, opportunities remain to improve the presentation of this information, especially in graphical form. The `factorplot` command and accompanying methods discussed in this article graphically and numerically present results of hypothesis tests related to pairwise comparisons resulting from qualitative covariates in GLMs or coefficients in multinomial logistic regression models.

Introduction

The problem of presenting information about categorical covariates in generalized linear models is a relatively simple one. Nevertheless, it has received some attention in the recent literature. To be clear about the problem, consider the following linear model where y is the dependent variable and $G = \{1,2,\ldots,m\}$ is a categorical independent variable that can be represented in the regression model by $m-1$ dummy regressors, each one representing a different category of G. The reference category, of course, is omitted. Thus, the model looks as follows:

$$E(y_i) = \mu_i \tag{1}$$

$$g(\mu_i) = \beta_0 + \beta_1 D_{i1} + \beta_2 D_{i2} + \cdots + \beta_{m-1} D_{im-1} + \beta_m X_{i1} + \ldots + \beta_{m+k-1} X_{ik} + \varepsilon_i, \tag{2}$$

where $D_{i1} = 1$ if $G_i = 1$, $D_{i2} = 1$ if $G_i = 2$, etc. X_{ik} represent an arbitrary set of additional variables of any type. Here, each of the coefficients on the dummy regressors for G ($\beta_1,\ldots,\beta_{m-1}$) gives the difference in the conditional transformed mean of y between the category represented by the dummy regressor and the reference category, controlling for all of the other X_{ik}. However, the $m-1$ coefficients for the categories of G imply $\frac{m(m-1)}{2}$ simple contrasts representing every pairwise comparison between categories of G. Any single pairwise comparison of non-reference category coefficients can be conducted in a straightforward fashion. If the goal is to discern whether the conditional mean of y given $G = 1$ is different from the conditional mean of y given $G = 2$ holding all of the X variables constant, the quantity of interest is:

$$t = \frac{b_1 - b_2}{\sqrt{V(b_1 - b_2)}}, \tag{3}$$

where

$$V(b_1 - b_2) = V(b_1) + V(b_2) - 2V(b_1, b_2). \tag{4}$$

Thus, the calculation is not difficult, but calculating and presenting all of these differences can become cumbersome, especially as m gets large.[1] The problem comes not in the calculation of these quantities, but in the parsimonious presentation of this information that will allow users to evaluate any desired (simple) contrasts. Below, I discuss two extant methods used to present such information. Floating absolute risk first suggested by Easton et. al. (1991) and more rigorously justified by Firth and Menzes (2004); Menzes (1999); Plummer (2004) - a method of overcoming the reference category problem by calculating floating variances for all levels of a factor (including the reference category). These floating variances can be used to perform hypothesis tests or construct floating confidence intervals that facilitate the graphical comparison of different categories (i.e., [log-]relative risks). The multiple comparisons literature has traditionally been focused on finding the appropriate p-values to control either the family-wise error rate (e.g., Holm, 1979) or the false discovery rate (e.g., Benjamini and Hochberg, 1995) in a set of simultaneous hypothesis tests. Presentation of this information has either been in the form of line displays (e.g., Steel and Torrie, 1980) or compact letter displays (e.g., Gramm et al., 2007).

However, when simple contrasts are the only quantities of interest, neither method above is perfect. When floating/quasi-variances are presented, the user still has to evaluate a potentially large number of hypothesis tests by either relying on the overlap in the floating confidence intervals or by calculating the floating t-statistic. Either solution requires a good deal of cognitive energy on the part of the analyst or reader. Compact letter displays do well at identifying patterns of statistical significance, but are perhaps cumbersome to investigate when patterns of (in)significance are complicated. Below, I discuss a method that presents this information in a manner that will permit the immediate evaluation of all the

[1] Tools to carry out these computations already exist in the `multcomp` package in R (Hothorn et. al., 2008).

$m(m-1)/2$ hypothesis tests associated with simple contrasts. The method I propose can also calculate analytical standard errors that are not prone to the same potential inferential errors produced by floating variances. Alternatively, the user may specify point estimates (e.g., log relative risks) and floating variances to produce the same plots and tests. The plot method for the function presents a graphical depiction of all hypothesis tests of interest that do not require the analyst to make any judgement about the degree of overlap of two confidence intervals and the extent to which that overlap is evidence of statistical significance of the difference of estimates.

Solutions to the Reference Category Problem

There are a number of reasonable solutions to the reference category problem.[2] The first solution is to present all of the covariance information required to calculate t-statistics for contrasts of interest (i.e., the variance-covariance matrix of the estimators). This solution provides the reader with all necessary information to make inferences. However, it does not provide an easy way for all of these inferences to be presented. Another solution is to re-estimate the model with different reference categories in turn.[3] This method produces the correct inferential information, but it is inelegant. The modal response to the reference category problem is a failure to do anything to discover (or allow readers to investigate) the implied pairwise differences not captured by the estimated coefficients.

Easton et. al. (1991) proposed the idea of floating absolute risk (FAR) as a means for evaluating multiple comparisons in matched case-control studies. The idea was to provide sufficient information such that readers could perform multiple comparisons with estimates of floating absolute risk at the expense of presenting a single extra number for each binary variable representing a level of a categorical covariate (i.e., risk factor). Although Greenland et al. (1999) disagreed on terminology and on the utility of Easton's idea of a floating scale, they agreed on the utility presenting information that would permit users to easily make the right inferences about relative risks among any levels of a categorical risk factor. Both Firth and Menzes (2004); Plummer (2004) provided a more rigorous statistical foundation on which to build estimates of floating absolute risk (or as Firth and Menzes call them, quasi-variances). Firth and Menzes' method has been operationalized in R in the **qvcalc** package (Firth, 2010) and both the methods of Plummer as well as Greenland et al. have been operationalized in the `float()` and `ftrend()` functions, respectively, in the **Epi** package (Bendix et al., 2010). In general, these solutions allow sufficient information to be presented in a single column of a statistical table that makes valid, arbitrary multiple comparisons possible.

The measures of floating absolute risk are often used to create floating (or quasi-) confidence intervals.[4] Presenting these intervals allows the user to approximately evaluate hypothesis tests about any simple contrast. While the exact nature of these confidence intervals is somewhat controversial (for a discussion, see Easton and Peto (2000); Greenland et al. (1999, 2000)), all agree that confidence intervals can be profitably put around some quantity (either the log-relative risks versus the reference category or the floating trend) to display the uncertainty around these quantities and permit visual hypothesis tests.

However, this still require the analyst or reader to either evaluate the pairwise hypothesis tests based on the extent to which confidence intervals overlap or calculate the floating t-statistic for each desired contrast. If the former, readers must still engage in a cognitive task of position detection (Cleveland, 1985) and then make an inference based on the extent to which intervals overlap. As the distance between floating confidence intervals grows, this task becomes more difficult. Finally, as Easton et. al. (1991) suggests, floating variances are a "virtually sufficient" summary of the uncertainty relating to relative risks; however, they can produce erroneous inferences if the error rate is sufficiently high. Both Firth and Menzes (2004) and Plummer (2004) provide methods for calculating this error rate, which is often small relative to other sources of error in the model.

To put a finer point on the problem, consider the example below using data from Ornstein (1976). The model of interest is:

$$\begin{aligned} \text{Interlocks}_i &\sim \text{Poisson}(\mu_i) \\ \log(\mu_i) &= \beta_0 + \beta_1 \log_2(\text{Assets}_i) \\ &\quad + \boldsymbol{\gamma}\text{Sector}_{ij} + \boldsymbol{\theta}\text{Nation}_{im} \end{aligned} \tag{5}$$

where γ represents a set of coefficients on the $j = 9$ non-reference category dummy variables for the 10 sectors represented in the data and $\boldsymbol{\theta}$ is the set of coefficients for the $m = 3$ coefficients on the non-reference category dummy variables representing the four nations in the dataset. The goal is to determine which

[2] The problem here applies particularly to polytomous, unordered risk factors or covariates. The case of ordinal risk factors, where only the difference in adjacent categories is of interest, is a bit less troublesome and will not be dealt with here.

[3] In fact, this re-parameterization method could be used to deal with more complicated contrasts, too. For example, it could be used to deal with the problem proposed by Greenland et al. (1999) wherein they wanted to estimate the relative risk of being above particular category on birthweight.

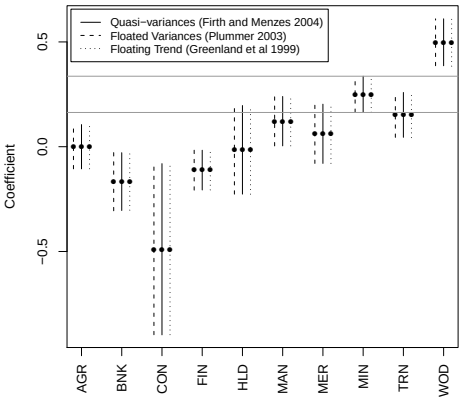
[4] Occasionally, quasi-variance estimates are negative, which provide the right inferences, but do not permit plotting of quasi-confidence intervals.

sectors (and/or nations) have significantly different conditional means of *Interlocks*. The quasi-variances can be presented along with the coefficients permitting hypothesis testing at the discretion of the reader. This approach is economical, but still requires the interested reader to make 27 pairwise hypothesis tests for sector and three pairwise hypothesis tests for nation, beyond those presented in the coefficient table.

The plot of the floating confidence intervals provides similar information, but readers are still required to make judgements about statistical significance that are occasionally difficult to justify. Consider Figure 1, which presents confidence intervals using the three different commands that produce floating variances R (R Development Core Team, 2010) `qvcalc()`, `float()` and `ftrend()` .[5] In the figure, the floating confidence interval for the mining sector overlaps four other floating confidence intervals and does not overlap the remaining five intervals.[6] Advice from Smith (1997) suggests that only confidence intervals not containing the point estimate against which the test is being done are significant. Here, all of the pairwise differences with the mining coefficient are significant because none of the point estimates are within the 95% confidence interval for mining. A more conservative strategy is to fail to reject null hypotheses where confidence intervals overlap and to reject otherwise. Using this criterion, the mining sector is different from five other coefficients - Agriculture, Banking, Construction, Finance and Wood. Browne (1979) shows that making inferences from confidence intervals requires a knowledge of the different sampling variances of the underlying random variables for which the confidence intervals have been constructed (i.e., the widths of the intervals matter); the decision does not rest solely on the extent to which the intervals overlap. While Browne's method may produce more appropriate inferences, it is hardly less work than producing the hypothesis tests directly. When the appropriate pairwise hypothesis tests are performed, without adjusting the *p*-values for multiple testing, it is clear that the mining coefficient is different from eight or seven coefficients, when using a one- or two-sided test, respectively.

Even if the evidence regarding the outcome of a hypothesis test from two confidence intervals is clear, there are other potential sources of error. Cleveland (1985) finds that detecting position along a common scale is one of the easiest tasks of graphical perception, but that discerning length is considerably more difficult. His experiments show that readers are prone to errors in even the easiest graphical perception tasks and the error rate is nearly twice as high when readers are asked to adjudicate the relative lengths of lines. Conducting hypothesis tests using confidence intervals is an endeavor rife with opportunities to make inferential errors.

Figure 1: Quasi-confidence Intervals for the Ornstein Model



Methods for calculating and presenting models with multiple simple contrasts have developed in the multiple testing literature as well. While the thrust of the literature mentioned above was dealing with the reference category problem directly, the multiple comparisons literature has placed greater focus on finding the appropriate *p*-values for a set of hypothesis tests rather than a single test. This can be accomplished through controlling the family-wise error rate (the probability of committing a Type I error on *any* of the tests in the set) or the false discovery rate (the proportion of falsely rejected hypotheses among those rejected). Chapter 2 of Bretz et. al. (2010) provides a brief, but informative discussion of these general concepts. While these are useful concepts, and the package discussed below permits users to adjust *p*-values in a number of ways to address these issues, I am more interested in how the multiple testing literature has developed around the presentation of multiple pairwise comparisons.

Gramm et al. (2007) discuss the two generally accepted methods for presenting multiple comparisons - the line display and the letter display. A line display (see for example, Steel and Torrie, 1980) prints a column where each row represents a single element in the multiple comparisons. In the example above, using the Ornstein data, these would be the names of the various sectors. Then, vertical lines are drawn connecting all values that are not significantly different from each other. This is a relatively simple display, but as shown generally by Piepho (2004) in this

[5]The figure below subtracts the arbitrary constant from the results of `ftrend()` to put all of these estimates on the same scale. I recognize that this is not what the authors had intended, but this should not lead to erroneous inferences in any event (Easton and Peto, 2000).

[6]Horizontal gray lines have been drawn at the smallest lower- and largest upper-bounds of the mining sector floating confidence intervals to facilitate comparison. Note that differences across the three methods in the upper bounds and lower bounds were in the third decimal place.

particular case, it is not always possible to faithfully represent all of the pairwise comparisons with connecting line segments. Figure 2(a) shows the line display for the Ornstein model above. Note that in the third line, there a discontinuity is required to properly depict all of the pairwise relationships. A compact letter display (Piepho, 2004) places a series of letters by each level of the categorical variable such that any two levels with the same letter are not significantly different from each other. These are more flexible than line displays, they can still be improved upon. Even though these displays do identify all pairwise significant relationships, they do not immediately identify the sign and size of the differences and what appear to be complicated patterns of significance may appear more simple with a different method of display.

An alternative method of presentation

A good solution to the reference category problem is one that permits the most efficient presentation and evaluation of a series of hypothesis tests relating to various (simple) factor contrasts. As discussed above, both the numerical presentation of floating variances and the visual presentation of floating confidence intervals are not maximally efficient on either dimension (presentation or evaluation) when the analyst desires information about the simple pairwise difference between coefficients related to the levels of a factor (i.e., simple contrasts). Similarly, I suggested that compact letter displays, though they present all of the appropriate information, are not maximally efficient at presenting the desired information graphically. As Chambers et al. (1983) and Cleveland (1985) suggest, one efficient way of presenting many pairwise relationships is through a scatterplot matrix or a generalized draftsman's display (a lower- or upper-triangular scatterplot matrix).[7] The important feature of a scatterplot matrix is the organization of pairwise displays in a common scale. Thus, a display that directly indicates the difference for the simple contrasts of interest would be superior to one that requires the user to make $(m(m-1))/2$ pairwise comparisons from m floating variances or confidence intervals.

The `factorplot` command in the library of the same name (version 1.0) for R computes all pairwise comparisons of coefficients relating to a factor; its print, summary and plot methods provide the user with a wealth of information regarding the nature of the differences in these coefficients. These functions overcome the problems suffered by previous methods as they present the results of pairwise hypothesis tests directly in a visually appealing manner.

The command calculates equation 3 for each simple contrast directly through a set of elementary matrix operations. First, $\boldsymbol{d}$, a $m \times \frac{m(m-1)}{2}$ matrix in which each column has one entry equal to positive one, one entry equal to negative one and all the remaining entries equal to zero is created. The positive and negative ones indicate the comparison being calculated. Using the coefficients for the desired factor covariate (call them $\boldsymbol{g}$, a row-vector of length m), I calculate $\Delta = \boldsymbol{g}\boldsymbol{d}$. Standard errors for each contrasts are calculated using the m rows and columns of the variance-covariance matrix of the estimators from the model (call this $V(g)$): $V(\Delta) = \boldsymbol{d}'\boldsymbol{V}(g)\boldsymbol{d}$. The Δ vector and the square root of the diagonal of $V(\Delta)$ (both of length $\frac{m(m-1)}{2}$) are then organized into $(m-1) \times (m-1)$ upper-triangular matrices where the rows refer to the first $m-1$ elements of $\boldsymbol{g}$ and the columns refer to the last $m-1$ elements of $\boldsymbol{g}$. The entries indicate the difference between the coefficient represented by the row and the coefficient represented by the column and its standard error.

The `factorplot2` command operates in a slightly different manner allowing the user to provide a set of estimates and a variance-covariance matrix for which pairwise comparisons are to be calculated. If the user provides a vector of floating or quasi-variances, the vector will be turned into a diagonal matrix and used in the calculations as described above. The plot, print and summary methods work in similar fashion for objects containing output from the `factorplot2` command.

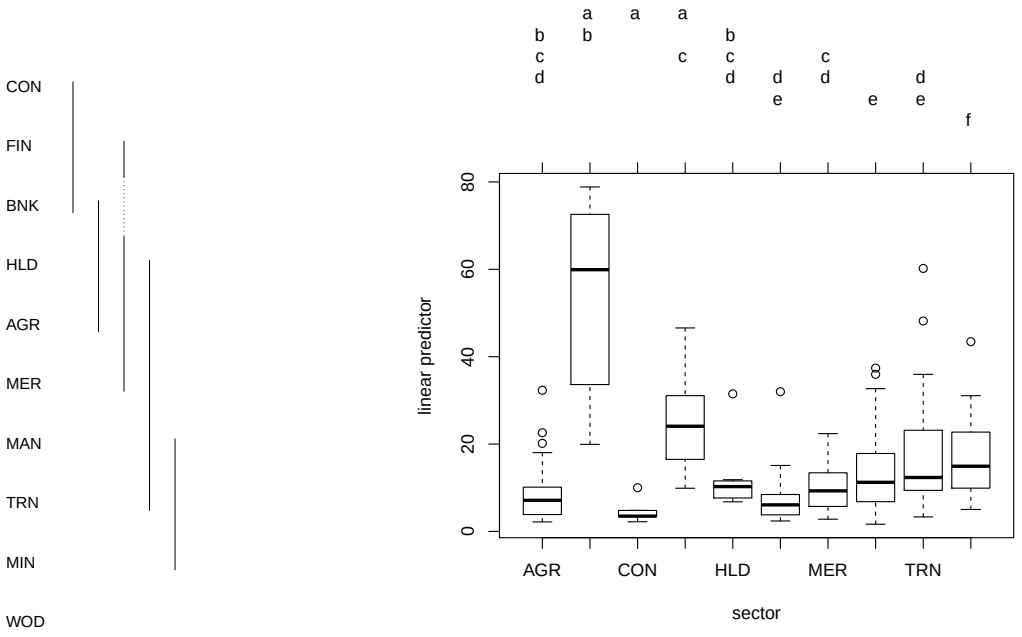
Example 1: Ornstein Data

The `factorplot` command has six arguments. The first two arguments, `obj` and `factor.variable` indicate the GLM object and the name of the factor for which comparisons are desired, respectively. The third argument, `pval`, allows the user to set the desired Type I error rate. The fourth argument, `two.sided` allows the user to specify whether the null hypothesis is tested against a one- or two-sided alternative with the latter as the default. The `order` argument sets the ordering of the coefficients, with three possibilities - 'natural', 'alph' and 'size'. The 'natural' option maintains the original ordering of the factor, the 'alph' option sorts them alphabetically and the 'size' option sorts in ascending order of the magnitude of the coefficient. The choices made here propagate through the plot, print and summary methods. Finally, the `adjust.method` argument allows users to adjust p-values to control either the family-wise error rate or the false discovery rate using the `p.adjust` function in the **stats** package.

The plot method for `factorplot` produces something akin to an upper-triangular scatterplot. The

[7] Cleveland (1985) makes the argument in favor of a full scatterplot matrix, but in this case, the information presented in the upper-triangle is sufficient as nothing new could be learned by examining the full square matrix.

Figure 2: Line and Letter Displays for Ornstein Model



(a) Line Display

(b) Letter Display

analogy is not perfect, but the idea is similar; each entry of the rows-by-columns display indicates the pairwise difference between coefficients. The statistical significance of these differences is indicated by three colors (one for significant-positive, one for significant-negative and one for insignificant differences).[8] The three colors can be controlled with the `polycol` argument and the text color within the polygons can be controlled with the `textcol` argument.[9] The plot method also allows the user to specify the number of characters with which to abbreviate the factor levels through the `abbrev.char` argument. Setting this to an arbitrarily high value will result in no abbreviation. Finally, the `trans` argument allows the user to impose a post-hypothesis-test transformation to the coefficient estimates. For example, if the underlying model is a logistic regression, tests will be done on the log-relative risks , but the relative risks could be plotted with `trans = "exp"`.[10] By default, the function prints legends identifying the colors and numbers; these can be turned on or off with the logical arguments `print.sig.leg` and `print.square.leg`, respectively. Figure 3 shows the display for the Ornstein model. The following code produces the result in the figure.

```
library(factorplot)
mod <- glm(interlocks ~ log2(assets) +
nation + sector, data=Ornstein,
family=poisson)
fp <- factorplot(mod, "sector",
    pval = 0.05,
    two.sided=TRUE,
    order="natural",
adjust.method="none")
plot(fp, abbrev.char=100)
```

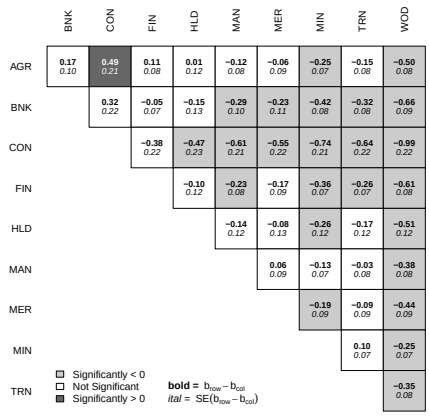
The print method for a `factorplot` object prints all of the pairwise differences, their accompanying analytical standard errors and (optionally adjusted) p-values. The user can specify the desired number of decimal places for rounding, with the `digits` argument. The `sig` argument is logical allowing the user to print all pairwise differences if `FALSE` and only significant differences when `TRUE`. The print method also permits the same `trans` argument as the plot method for objects of class `factorplot`. An example of the output from the print method is below. Here, twenty-five of the forty-five pairwise differences are statistically different from zero when.

[8] The choices made with respect to `adjust.method` persist through the plot, print and summary methods.

[9] Note, that polygons can be made to appear without text by setting `polycol` equal to `textcol`.

[10] After the hypothesis tests are done, a matrix named `r.bdiff` holds the coefficient differences. The transformation is done as follows: `do.call(trans, list(r.bdiff))`, so only transformation amenable to this procedure will work.

Figure 3: Plotted factorplot object for Ornstein model




```
print(fp, sig=T)
          Difference    SE p.val
AGR - CON      0.489 0.213 0.023
CON - HLD     -0.474 0.235 0.045
BNK - MAN     -0.288 0.102 0.005
CON - MAN     -0.611 0.215 0.005
FIN - MAN     -0.233 0.082 0.005
BNK - MER     -0.228 0.106 0.032
CON - MER     -0.551 0.220 0.013
AGR - MIN     -0.250 0.069 0.000
BNK - MIN     -0.416 0.084 0.000
CON - MIN     -0.739 0.210 0.001
FIN - MIN     -0.361 0.067 0.000
HLD - MIN     -0.265 0.118 0.026
MER - MIN     -0.188 0.085 0.029
BNK - TRN     -0.318 0.082 0.000
CON - TRN     -0.641 0.217 0.004
FIN - TRN     -0.263 0.070 0.000
AGR - WOD     -0.498 0.076 0.000
BNK - WOD     -0.665 0.095 0.000
CON - WOD     -0.988 0.215 0.000
FIN - WOD     -0.610 0.077 0.000
HLD - WOD     -0.513 0.121 0.000
MAN - WOD     -0.376 0.080 0.000
MER - WOD     -0.437 0.090 0.000
MIN - WOD     -0.248 0.072 0.001
TRN - WOD     -0.346 0.081 0.000
```

The summary method for `factorplot` prints the number of coefficients that are significantly smaller than the one of interest and the number of coefficients larger than the one of interest for each level of the factor. While this is not a common means of presenting the results, this does nicely summarize the extent of significant differences among the coefficients. Below is an example of printout from the summary method. It is easy to see that the wood industry (WOD) has the highest conditional means as it is significantly bigger than all of other categories. It is also easy to see that the construction industry (CON) has one of the smallest conditional means as it is significantly smaller than seven of the other categories and not significantly bigger than any.

```
summary(fp)
    sig+ sig- insig
AGR    1    2     6
BNK    0    5     4
CON    0    7     2
FIN    0    4     5
HLD    1    2     6
MAN    3    1     5
MER    2    2     5
MIN    6    1     2
TRN    3    1     5
WOD    9    0     0
```

Together, the `factorplot` command and its associated print, plot and summary methods provide a wealth of information including direct hypothesis tests using analytical standard errors for the simple contrasts most commonly desired in (G)LMs.

Example 2: *H. pylori* and Gastric Precancerous Lesions

Plummer et al. (2007) were interested in discerning the extent to which infection with *H. pylori* containing the cytotoxin-associated (cagA) gene increased the severity of gastric precancerous lesions. They found that cagA+ patients had increased risks of more severe lesions while cagA- patients were only at significantly higher risk (than their uninfected counterparts) of chronic gastritis. Table 1 summarizes the results of the relative risk of the various types of gastric lesions versus the baseline of normal or superficial gastritis. The presence of floating standard errors makes it relatively easy to construct floating confidence intervals and calculate hypothesis tests.

The `factorplot2()` command allows the user to supply a vector of point estimates and (floating) variances rather than an estimated model object. This function will be particularly useful for those scholars in epidemiology, where floating standard errors are more routinely presented. From the third column of Table 1, it would seem that that there are two significant pairwise differences, where floating confidence intervals do not overlap. Namely, *H. pylori* cagA- seems to raise the risk of chronic gastritis relative to Intestinal metaplasia I and the reference group of normal and superficial gastritis. The sixth column of Table 1 indicates that *H. pylori* cagA+ significantly increases the risk of all other forms of gastritis relative to normal and superficial. Further, it appears that the risk of chronic gastritis, chronic atrophic gastritis and intestinal metaplasia I are not significantly different from each other and similarly the risks of intestinal metaplasia II and III and dysplasia are not statistically different from each other. It does appear that the risk of any of the outcomes in the latter group of three is statistically higher than the risk of the outcomes in the former group of three, though.[11]. Using `factorplot2()`, I can provide a more precise test of the differences. Below is an example of how

[11] Strictly speaking, the difference between intestinal metaplasia I and II appears to be just insignificant as there is a small overlap in the intervals

Table 1: Results from Plummer et al. (2007)

	cagA-			cagA+		
	OR	FSE	95% FCI	OR	FSE	95% FCI
Normal and superficial gastritis	1.00	0.242	(0.62, 1.61)	1.00	0.320	(0.53, 1.87)
Chronic gastritis	2.12	0.096	(1.76, 2.56)	4.33	0.101	(3.55, 5.28)
Chronic atrophic gastritis	1.44	0.156	(1.06, 1.96)	3.89	0.160	(2.84, 5.32)
Intestinal metaplasia I	1.31	0.140	(1.00, 1.72)	4.14	0.141	(3.14, 5.46)
Intestinal metaplasia II	1.44	0.380	(0.68, 3.03)	10.8	0.349	(5.45, 21.40)
Intestinal metaplasia III	1.46	0.484	(0.57, 3.77)	21.9	0.431	(9.41, 50.97)
Dysplasia	0.90	0.375	(0.43, 1.88)	15.5	0.311	(8.43, 28.51)

OR = odds ratio
FSE = floating standard error
FCI = floating confidence interval (calculated by the author, not presented in Plummer et al. [2007])
Adapted from Figure 1 in Plummer et al. (2007, 1331).

the results could be used to in conjunction with the `factorplot` suite of functions.

```
est1 <- log(c(1.00,2.12,1.44,1.31,1.44,
     1.46,0.90))
var1 <- c(0.242,0.096,0.156,0.140,
     0.380,0.484,0.375)^2
est2 <- log(c(1.00,4.33,3.89,4.14,10.8,
     21.9,15.5))
var2 <- c(0.320,0.101,0.160,0.141,
     0.349,0.431,0.311)^2
resdf <- 48+16+27+532+346+144+144+124+
     58+166+162+75+24+53+10+15+61+6+
     18+90+12-18
names(est1) <- names(est2) <- c(
     "Normal & superficial gastritis",
     "Chronic gastritis",
     "Chronic atrophic gastritits",
     "Intestinal metaplasia I",
     "Intestinal metaplasia II",
     "Intestinal metaplasia III",
     "Dysplasia")

plummer_fp1 <- factorplot2(est1, var1, resdf,
    adjust.method="none")
plummer_fp2 <- factorplot2(est2, var2, resdf,
    adjust.method="none")
plot(plummer_fp1, trans="exp",
    abbrev.char = 100)
plot(plummer_fp2, trans="exp",
    abbrev.char = 100)
```

The plots are displayed in Figure 4. Both of the differences that appeared significant are according to the factorplot as are two of the differences that exhibited minimal overlap in the confidence intervals. The differences in the risk of chronic gastritis and chronic atrophic gastritis or dysplasia are also significant according to the floating t-statistics calculated by `factorplot2`. The factorplot on the right indicates that there are no significant differences among the second through fourth diagnoses and the fifth through seventh diagnoses. However, all other pairwise differences are significant. Again, relying strictly on the overlap in confidence intervals suggests that the difference between the risk of intestinal metaplasia I and II (for cagA+) is not significant, though evaluating the floated t-statistic indicates otherwise.

Example 3: Vote Choice in France

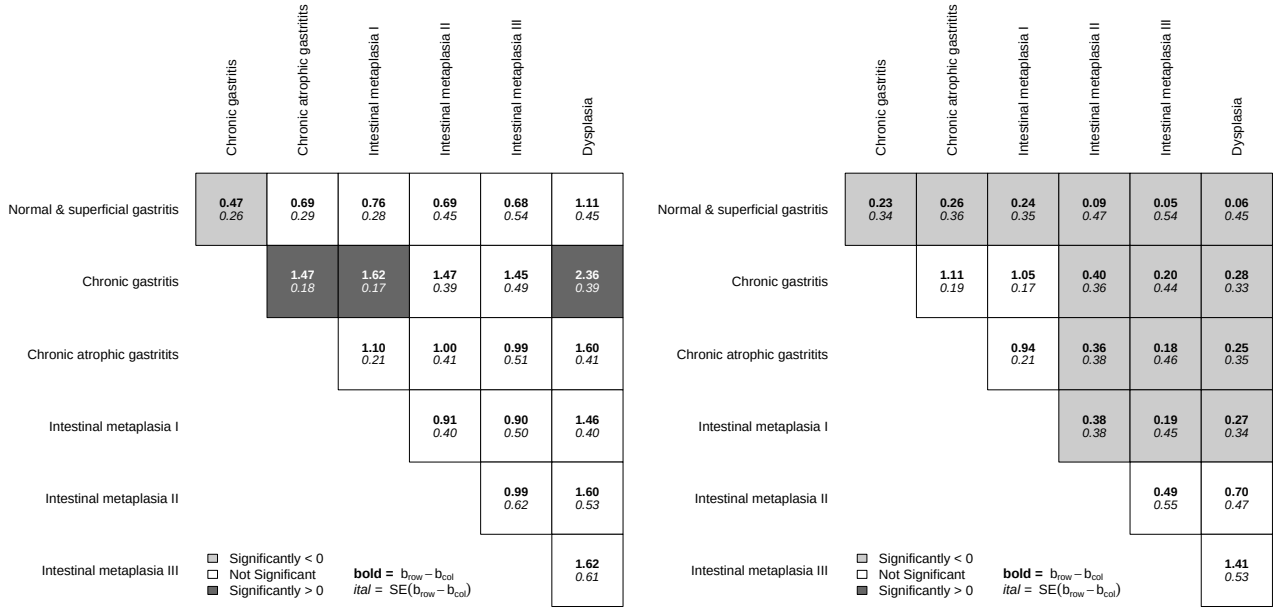
The `factorplot2` command is versatile enough to be used whenever pairwise comparisons need to be made across a set of estimates with a specified variance-covariance matrix. One such situation is with multinomial logistic regression coefficients. The coefficient table presents a specific set of pairwise comparisons - namely those indicating the relationship of each variable to the choice of voting for each non-reference party versus the reference party. However, other comparisons may be interesting or useful.

In the example below, I estimate a MNL model of vote choice (`vote`) on a number of standard controls: retrospective national economic evaluations (`retnat`), self-placement on the left-right ideological continuum (`lrself`), gender (`male`) and age (`age`). By saving the coefficients in such a way that their entires match up with the appropriate variance and covariances, these objects can be used as input to `factorplot2`. The results can be printed, summarized or plotted just as above. While there are a number of steps in the example below, they are easily changed to produce similar plots or summaries for other variables.

```
library(nnet)
data(france94)
france.mod <- multinom(vote ~ retnat +
    lrself + male + age, data=france)

# save coefficients and variances
b <- coefficients(france.mod)
```

Figure 4: Results from Plummer et al. (2007) Presented as factorplots



(a) cagA- (b) cagA+

```
v <- vcov(france.mod)

# create coefficient names
nb <- outer(rownames(b), colnames(b),
    paste, sep=":")

# save names in same order as varinaces
b.vec <- c(t(b))
names(b.vec) <- c(t(nb))

# find the age coefficients
age.ind <- grep("age",
    names(b.vec))

# extract age coefs and variances
age.b <- b.vec[age.ind]
age.v <- v[age.ind, age.ind]

# include coefs and vars that were
# set to 0 for identification
age.b <- c("PCF:age" = 0, age.b)
age.v <- rbind(0, cbind(0, age.v))

# drop ':age' extensions and rename
names(age.b) <- gsub(":age", "",
    names(age.b), fixed=T)
rownames(age.v) <- names(age.b)
colnames(age.v) <- names(age.b)

# calculate residual df
res.df <- with(france.mod,
    length(fitted.values) - edf)
```

```
# run factorplot2 command
fp3 <- factorplot2(age.b, age.v,
    res.df, adjust.method="none")

# create plot
plot(fp3)
```

Figure 5: Plotted factorplot object for Age from Multinomial Logit model

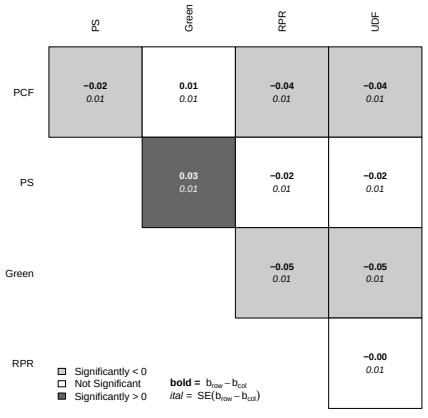


Figure 5 shows that as people get older, they are more likely to vote for RPR or UDF than the Greens or Communists (PCF) and more likely to vote for the Socialists (PS) than the Greens. If one is interested in whether variables have significant effects on vote choice, all pairwise comparisons should be considered. `factorplot` makes it easy for users to appropriately evaluate all relevant pairwise comparisons.

Conclusion

Easton's (1991) contribution of floating absolute risk has been influential, especially in epidemiology and medicine, allowing researchers to present easily information that permits the reader to make any pairwise comparison among the different levels of a risk factor. Firth and Menzes (2004); Menzes (1999) and Plummer (2004) have provided not only a rigorous, model-based foundation for this idea, but have also provided software that easily produces these quantities for a wide array of statistical models. I argue that while these quantities are interesting and useful, floating confidence intervals, which are often provided ostensibly to permit hypothesis testing can be imprecise and potentially misleading, as regards hypothesis testing. Compact letter displays (Piepho, 2004) are a step in the right direction, but I argue that they can still be improved upon in terms of graphically presenting information of interest to many researchers. In the common situation wherein one is interested in simple contrasts, the `factorplot` and `factorplot2` commands and their associated print, plot and summary methods discussed above provide much greater transparency with respect to the presentation and evaluation of hypothesis tests than floating absolute risk or quasi-variance estimates. The visual presentation of direct hypothesis tests requires much less effort to adjudicate significance and uncover patterns in the results than other methods, including compact letter displays. While the calculation of these hypothesis tests is not novel, the methods of presenting and summarizing the information represent a significant advance over the previously available general solutions available in R.

Bibliography

C. Bendix, M. Plummer, E. Laara and M. Hills. *Epi: A package for statistical analysis in epidemiology*, 2010. URL `http://CRAN.R-project.org/packag=Epi`. R package version 1.1.15.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57:, 289–300.

F. Bretz, T. Hothorn and P. Westfall. *Multiple Comparisons Using R*. CRC Press, Boca Raton, FL, 2010.

R. H. Browne. Visual assessment of the significance of a mean difference. *Biometrics*, 35(3):657–665, 1979.

J.M. Chambers, W.S. Cleveland, B. Kleiner, and P.A. Tukey. *Graphical Methods for Data Analysis*. Wadsworth & Brooks/Cole Publishing Company, Pacific Grove, CA, 1983.

W. Cleveland. *Elements of Graphing Data*. Wadsworth, Inc., Monterey, CA, 1985.

D.F. Easton, J. Peto and G.A.G. Babiker Floating absolute risk: An alternative to relative risk in survival and case control analysis avoiding an arbitrary reference group. *Statistics in Medicine* 10: 1025–1035, 1991.

D.F. Easton and J. Peto. Re: "Presenting statistical uncertainty in trends and dose-response relationships" (letter). *American Journal of Epidemiology*, 152:393, 2000.

D. Firth. Overcoming the reference category problem in the presentation of statistical models. *Sociological Methodology*, 33:1–18, 2003.

D. Firth. *qvcalc: Quasi variances for factor effects in statistical models*, 2010. URL `http://CRAN.R-project.org/package=qvcalc`. R package version 0.8-7.

D. Firth and R.X. Menzes. Quasi-variances. *Biometrika*, 91(1):65–80, 2004.

J. Gramm, J. Guo, F. Hüffner, R. Niedermeier, H. Piepho and R. Schmid. Algorithms for Compact Letter Displays: Comparison and Evaluation. *Computational Statistics and Data Analysis* 52(12): 725–736, 2007.

S. Greenland, K.B. Michels, J.M. Robins, C. Poole and W.C. Willett. Presenting statistical uncertainty in trends and dose-response relations. *American Journal of Epidemiology* 149(12):1077–1086, 1999.

S. Greenland, K.B. Michels, C. Poole and W.C. Willett. Four of the authors reply [Re: "Presenting statistical uncertainty in trends and dose-response relationships"] (letter). *American Journal of Epidemiology*, 152:394, 2000.

S. Holm A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 6: 65–70, 1979.

T. Hothorn, F. Bretz and P. Westfall. Simultaneous Inference in General Parametric Models. *Biometrical Journal*, 50:346–363, 2008.

J.C. Hsu. *Multiple Comparisons*. Chapman & Hall, London, 1996.

R.X. Menzes. *More Useful Standard Errors for Group and Factor Effects in Generalized Linear Models*. PhD thesis, Department of Statistics, University of Oxford, 1999.

M. Ornstein. The boards and executives of the largest canadian corporations. *Canadian Journal of Sociology*, 1:411–437, 1976.

H.P. Piepho. An Algorithm for a Letter-based Representation of All Pairwise Comparisons. *Journal of Computational and Graphical Statistics*, 13:456–466.

M. Plummer Improved estimates of floating absolute risk. *Statistics in Medicine* 23:93–104, 2004.

M. Plummer, L. van Doorn, S. Franceschi, B. Kleter, F. Canzian, J. Vivas, G. Lopez, D. Colin, N. Muñoz and I. Kato *Helicobacter pylori* cytotoxin-associated genotype and gastric precancerous lesions. *Journal of the National Cancer Institute*, 99:1328–1334, 2007.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2010. URL http://www.R-project.org. ISBN 3-900051-07-0.

J.P. Shaffer Multiple Hypothesis Testing. *Annual Review of Psychology* 46, 561–584.

R.W. Smith. Visual hypothesis testing with confidence intervals. *Proceedings of the 22nd SAS User's Group*, 1252–1257, 1997.

R.G.D. Steel and J.H. Torrie. *Principles and Procedures of Statistics.* McGraw-Hill, New York, 1980.

David A. Armstrong II
University of Wisconsin - Milwaukee
Department of Political Science
P.O. Box 413
Milwaukee, WI 53201
United States of America
armstrod@uwm.edu