Most studies addressing lexical processing make use of factorial designs. For many researchers in this field of inquiry, a real experiment is a factorial experiment. Methods such as regression and factor analysis would not allow for hypothesis testing and would not contribute substantially to the advancement of scientific knowledge. Their use would be restricted to exploratory studies at best. This paper is an apology coming to the defense of regression designs for experiments including lexical distributional variables as predictors.

In studies of the mental lexicon, we often are dealing with two kinds of predictors, to which I will refer as TREATMENTS and COVARIATES. Stimulus-onset asynchrony (SOA) is an example of a treatment. If we want to study the effect of a long versus a short SOA, it makes sense to choose sensible values, say 200 ms versus 50 ms, and to run experiments with these two settings. If the researcher knows that the effect of SOA is linear, and that it can be administered independently of the intrinsic properties of the items, then the optimal design testing for an effect of SOA is factorial. One would lose power by using a regression design testing for an effect at a sequence of SOA intervals, say 50, 60, 70, ..., 200 ms. This advantage of sampling at the extremes is well-known (see, e.g., Crawley, 2002, p. 67): the further apart the values of SOA are, the larger the corresponding sum of squares, and the smaller the standard error for the slope.

The advantage of designs with maximal contrasts for treatment predictors is often assumed to carry over to the study of lexical covariates such as frequency, length, neighborhood density, etc. In order to test for an effect of frequency, the traditional wisdom advises us to create a data set with very high-frequency words on the one hand, and very low-frequency words on the other hand. The problem that one runs into very quickly is that the set of high-frequency words will comprise short words with many neighbors, and that the low-frequency words will be long words with few neighbors. The massive correlations characterizing lexical properties create the problem that an effect of frequency could just as well be an effect of length or an effect of neighborhood density, or any combination of these variables. The traditional solution is to create a factorial contrast for frequency, while matching for the other predictors. This can be done by hand, or with the help of Maarten van Casteren’s mix program (Van Casteren and Davis, 2006). The aim of this contribution is to illustrate, by means of some simple simulations, that this matching process leads to a severe loss of power (following up on, e.g., Cohen, 1983; MacCallum et al., 2002).

In all the simulations to follow, the dependent variable (RT) is a function of two numerical predictors, $X_1$ (this could be log frequency, or the word’s imageability) and $X_2$ (this could be number of orthographic neighbors, or word length), which both follow a standard normal
Table 1: Correlations between five covariates for 2284 monomorphemic English nouns and verbs in the study of Baayen et al. (2006).

<table>
<thead>
<tr>
<th>Written Frequency</th>
<th>Family Size</th>
<th>N-Count</th>
<th>Familiarity</th>
<th>Length (in letters)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Written Frequency</td>
<td>1.00</td>
<td>0.66</td>
<td>0.10</td>
<td>0.79</td>
</tr>
<tr>
<td>Family Size</td>
<td>0.66</td>
<td>1.00</td>
<td>0.17</td>
<td>0.59</td>
</tr>
<tr>
<td>N-Count</td>
<td>0.10</td>
<td>0.17</td>
<td>1.00</td>
<td>0.10</td>
</tr>
<tr>
<td>Familiarity</td>
<td>0.79</td>
<td>0.59</td>
<td>0.10</td>
<td>1.00</td>
</tr>
<tr>
<td>Length (in letters)</td>
<td>-0.07</td>
<td>-0.12</td>
<td>-0.63</td>
<td>-0.08</td>
</tr>
</tbody>
</table>

The extent to which pairs of covariates correlate varies substantially, as illustrated in Table 1 for English monomorphemic nouns and verbs. Across simulations, I therefore varied the extent to which $X_1$ and $X_2$ correlate, with as smallest correlation $r = 0.2$, as medium correlation $r = 0.4$, and as largest correlation $r = 0.6$. The tighter this correlation, the more difficult it is to create a contrast in $X_1$ while matching in the mean for $X_2$.

The simulated RTs are defined in terms of $X_1$, $X_2$, with varying degrees of by-observation noise $\epsilon$ (with standard deviations ranging from 15 to 100) as follows:

\[
RT = 600 - 4X_1 - 4X_2 + \epsilon
\]  
\[
RT = 600 - 4X_1 - 4X_2 - 5X_1 \times X_2 + \epsilon
\]  
\[
RT = 600 - 1X_1 - 4X_2 + 6X_1 \times X_1 + \epsilon
\]

The corresponding regression surfaces (for random samples of data points) are shown in Figure 1. Contour lines connect points of the regression surface with the same simulated RT. In the left panel, contour lines are 5ms apart, in the central and right panels, they are 20 and 10 ms apart respectively.

In the left panel of Figure 1, $X_1$ and $X_2$ are both facilitatory, and do not interact, as can be seen from the parallel contour lines. In the central panel, the two predictors enter into a multiplicative interaction (cf., e.g., Kuperman et al., 2008, 2009, for examples from eye-tracking studies). RTs are longer towards the upper left and lower right corners, they are shorter towards the lower left and upper right corners, and intermediate in the center. This interaction is the analogue of the familiar X-shaped cross-over interaction for two factorial predictors. In the right panel, the effect of $X_1$ is U-shaped, but independent of the effect of $X_2$ (cf., e.g., Bien et al., 2005; Tabak et al., 2010, for U-shaped effects of frequency). For any given value of $X_2$, RTs first decrease and then increase. The regression surface has a vertical trough in the center which becomes deeper for greater values of $X_2$.

For each of the models (1)–(3), for each combination of $r$ (the correlation of $X_1$ and $X_2$), and for each level of by-observation noise $\epsilon$, 100 simulated data sets were created. For each data set, an attempt at matching for $X_2$ while factorially contrasting for high versus low $X_1$ was carried out. If no good matching was obtained, the simulated data set was discarded. Matching was accepted as satisfactory when a $t$-test did not detect a significant difference for $\alpha = 0.2$. In Figure 1, the data points selected for the factorial design are encircled.
For the data generated according to (1), a regression model with $X_1$ and $X_2$ as predictors was compared to a factorial model with $X_2$ as covariate. In the notation of $\mathbb{R}$,

\[
RT \sim X_1 + X_2 \\
RT \sim F + X_2
\]

with $F$ denoting the factorial contrast in $X_1$. For the data generated by (2), which introduces an interaction, the statistical models examined include the interaction. For the factorial data sets, this amounts to an analysis of covariance.

\[
RT \sim X_1 \times X_2 \\
RT \sim F \times X_2
\]

The data sets implementing a U-shaped effect of $X_1$ were modeled with a quadratic polynomial when using regression. The factorial design cannot test for nonlinearities, so only an effect of the factorial contrast is examined.

\[
RT \sim \text{poly}(X_1, 2, \text{raw=TRUE}) + X_2 \\
RT \sim F + X_2
\]

Figure 2 graphs the proportions of data sets, generated according to model (1), for which a significant effect was observed. Each panel summarizes the results for a different correlation between $X_1$ and $X_2$, for six different amounts of by-observation noise, for analysis of variance (left) and regression (right).

Across all panels, we find that the regression analysis outperforms the analysis of variance with a factorial contrast (almost all lines connecting the observed proportions of significant effects for analysis of variance and regression have a non-zero, non-negligible positive slope). It is only when $X_1$ and $X_2$ are weakly correlated, with little by-observation noise, that
Figure 2: Model (1), with $X_1$ and $X_2$ as independent predictors. Proportion of significant contrast coefficients for $X_1$ when dichotomized and matched for $X_2$ in analysis of variance, and proportion of significant slopes in a regression analysis with $X_2$ as covariate, for varying degrees of correlation between $X_1$ and $X_2$ ($r$), and varying degrees of by observation noise (standard error). Across the board, regression outperforms analysis of variance.
Figure 3: Model (2), including an interaction of $X_1$ by $X_2$. Proportion of significant contrast coefficients for $X_1$ when dichotomized and matched for $X_2$ in analysis of variance, and proportion of significant slopes in a regression analysis with $X_2$ as covariate, for varying degrees of correlation between $X_1$ and $X_2$ ($r$), and varying degrees of by observation noise (standard error). Upper panels: coefficients for the main effect of $X_1$, lower panels: coefficients for the interaction of $X_1$ and $X_2$. Across the board, regression outperforms analysis of variance.

the two methods perform equivalently at ceiling, as can be seen in the leftmost panel. As the correlation between $X_1$ and $X_2$ increases, the advantages of regression become more pronounced. The regression analyses can make use of all the data, affording it more power than the analyses of variance, which are limited by the shackles of the factorial contrast to smaller data sets, especially for larger $r$.

When $X_1$ and $X_2$ enter into an interaction (as illustrated in the central panel of Figure 1), analysis of variance can be exchanged for analysis of covariance, allowing the slope of $X_2$ to vary for the high versus low factor levels created for $X_1$. The upper panels of Figure 3 summarize the power of detecting a factorial contrast (left points), and the power of detecting the main effect of $X_1$ (right points). The lower panels present the proportion of simulation runs in which the contrast in the slope of $X_2$ is detected by the analysis of variance, and for the regression the proportion of runs in which the coefficient for the interaction term is
Figure 4: Model (3): the effect of $X_1$ is U-shaped. Proportion of significant contrast coefficients for $X_1$ when dichotomized and matched for $X_2$ in analysis of variance (upper panels), and proportion of significant linear (upper panels) and quadratic slopes (lower panels) in a regression analysis with $X_2$ as covariate, for varying degrees of correlation between $X_1$ and $X_2$ ($r$), and varying degrees of by observation noise (standard error). Across the board, regression outperforms analysis of variance.

Detected to be significant. Again, we find that regression substantially outperforms analysis of variance.

Figure 4 contrasts the performance of analysis of variance and regression for the case that $X_1$ has a U-shaped effect (see the right panel of Figure 1). Dichotomization makes it impossible to trace the non-linearity. However, as long as the linear coefficient of model (3) is not zero, it still can detect a difference between the two factor levels. Nevertheless, regression is much better at detecting (and estimating) the significance of the linear and quadratic terms of $X_1$.

What these simulation studies show is an unambiguous advantage to the use of regression compared to analysis of variance based on factor dichotomizing $X_1$. This advantage presents itself already for the simplest case in which the two predictors do not enter into an interaction. For more complex data sets in which the predictors interact, or in which one or
more predictors are nonlinear, regression offers the analyst the best opportunity to detect such interactions and nonlinearities.

In the present simulations, only one predictor had to be controlled during dichotomization. When many predictors have to be controlled for simultaneously, it becomes increasingly difficult to obtain properly matched materials. Indeed, many years ago, Anne Cutler pointed out that making up materials is a confounded nuisance (Cutler, 1981). Making up materials is indeed a nuisance when analysis of variance is the only statistical tool at one’s disposal. The larger the set of covariates, and the greater the correlational structure characterizing these predictors, the smaller one’s data set becomes. Sometimes, proper matching is not possible. In such a case, researchers may decide to relax their criteria for matching, or allow high and low subsets to have overlapping distributions on $X_1$. Alternatively, an interesting research question may rest unexplored even though regression would have allowed interesting insights to be obtained.

Another major concern is that the matching procedure violates the basic principle that one’s observations should be sampled randomly. An experiment with 10 matched words in each of the four cells of a dichotomized 2x2 factorial is an experiment about 40 very specific words, and does not allow generalization to the population of words. In the statistical analysis, such words should be entered as a fixed-effect factor, and not as a random-effect factor.

A related concern is that if the regression surface characterizing the population of words in which we are interested is characterized by complex non-linearities such as illustrated in Figure 1 (see Baayen et al., 2010, for more complex regression surfaces), then dichotomization amounts to taking snapshots of very restricted areas of this complex surface that (i) are not representative, and (ii) may not be easily replicable, leading to contradictory reports in the experimental literature.

It is often argued that a factorial design with a true treatment variable affords causal inference, whereas regression would only establish correlations, without allowing causal inference. For instance, if no priming effect is observed for a 3 ms prime duration, whereas a 40 ms priming effect is observed for a 60 ms prime duration, it makes sense to argue that increasing the prime duration causes a priming effect to emerge. Of course, the strength of this causal relation depends on how confident we are that when we change prime duration, there are no other variables that change as well. For example, the researcher should be 100% sure that the longer prime duration does not lead to a more challenging experiment that better engages the subjects. Good treatment variables have the property that we cannot easily conceive of other sensible confounding ‘hidden’ variables that might be involved.

Dichotomization is an attempt to create a treatment variable out of a correlational variable, supposedly allowing causal inference where otherwise causal inference would not be possible. However, regression implements matching mathematically, allowing inferences about the partial effect of one predictor on the dependent variable while other predictors are ‘held constant’. The present simulations show that mathematical matching through regression is much more effective than a-priori matching procedures preparing for analysis of variance. To reduce the likelihood that the partial effect observed for a given variable is not due to some
hidden variable lurking in the correlational structure of the data, it is essential to bring as many relevant predictors as possible into one’s regression model.

In summary, analysis of variance is appropriate for genuine treatment variables (SOA, prime duration, target presentation duration, etc.), provided that there is good reason to assume that the effect of the treatment variable is linear and unconfounded with other hidden variables. For predictors that are part of a complex correlational structure, dichotomization almost always leads to a loss of statistical power. For such predictors, a ‘real’ experiment is not a factorial experiment but a regression experiment.

References


