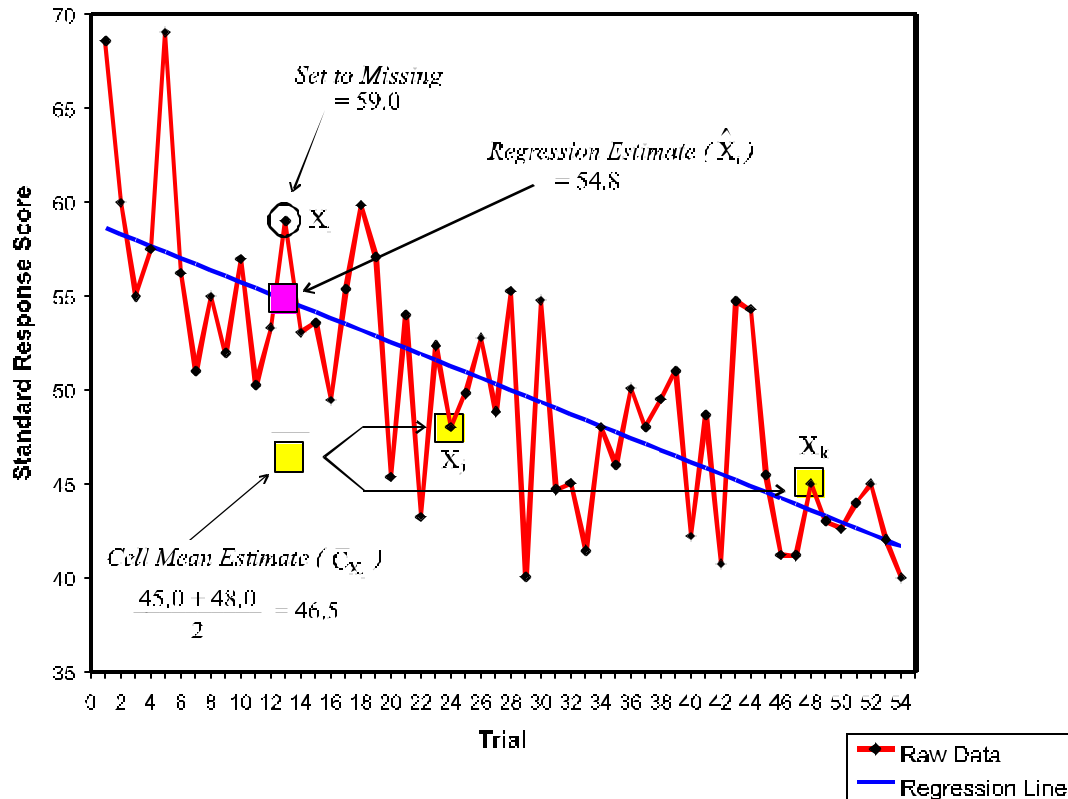


Estimation of missing data in psychophysiological research: Habituation should not be ignored.
J. Curtin & Christopher J. Patrick.

Frequently in psychophysiological experiments one or more data points are missing for many study participants. In within-subject designs, mean scores for each experimental condition (cell) are typically based on several trials, but no special method of handling missing data is routinely employed. Instead, cell means are simply derived from the remaining available data. This method--equivalent to estimating missing scores using the mean of other trials in that cell--would be adequate if all trials in the cell were expected to be comparable. However, this is often not the case. Many psychophysiological measures (e.g., startle blink, SCR) exhibit pronounced habituation, with the magnitude of scores decreasing substantially across trials. Thus, the accuracy of a cell mean estimate will vary depending upon the serial position of the missing data, with serious mis-estimates occurring for missing scores at either end of the habituation function.

An alternate regression-based approach to estimating missing data is proposed. This method uses trial as a predictor variable in a general linear model to estimate missing scores. Thus, it accounts for habituation by incorporating the serial position of the missing scores into the estimate. In the present study, a Monte Carlo analysis was conducted to compare the accuracy of the cell mean and regression methods in estimating startle magnitude for nine subjects in a slide viewing paradigm. The standard error of estimate (SEE), an index of the average difference between the observed and estimated scores, was computed for both methods across varying quantities of missing data.

Regression versus Cell Mean Estimation Methods

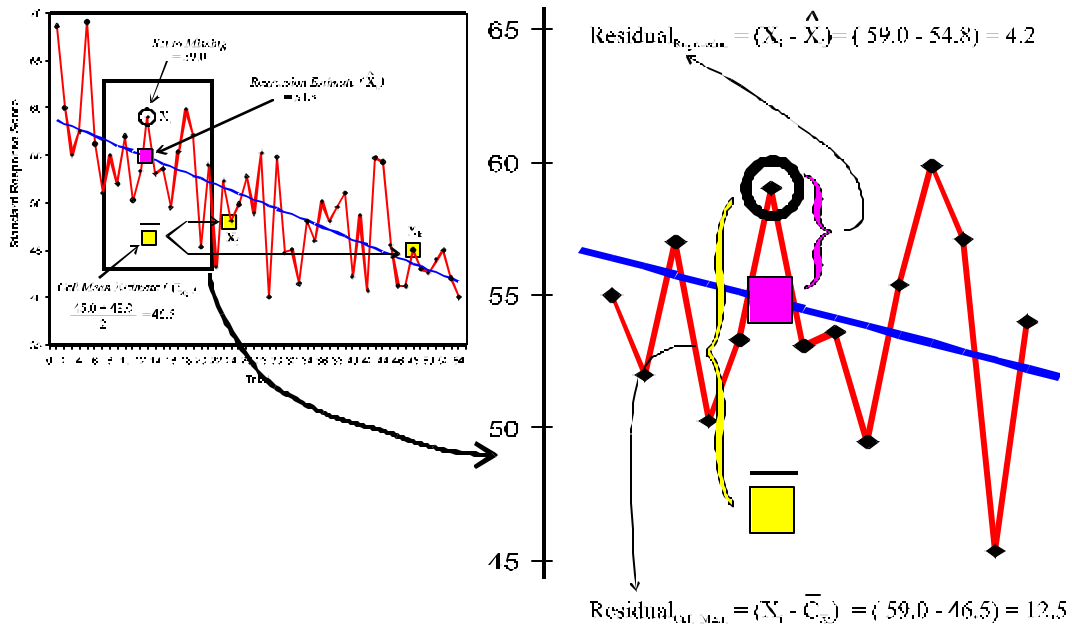


The above figure depicts the computations involved in estimating a single data point (X_i), arbitrarily set to missing, for a subject using Cell Mean and Regression methods. The raw data for this subject are plotted in red, and the best fitting regression line for trial as a predictor is plotted in blue.

If no explicit estimation procedure is used to predict the value of X_i , it is equivalent to substituting the Cell Mean (C_{X_i}) as an estimate. For the cell containing X_i , X_j and X_k are the two non-missing scores and thus the estimate of X_i is the mean of these two scores. However, as depicted above, if the dependent measure shows systematic habituation over trials, the cell mean method may provide a poor estimate of the missing score.

A more precise approach is to use Trial in a regression-based prediction equation to estimate X_i . The best fitting regression line for all non-missing data is calculated, and the estimate for the missing score is based on the predicted value (X_i) for that corresponding trial.

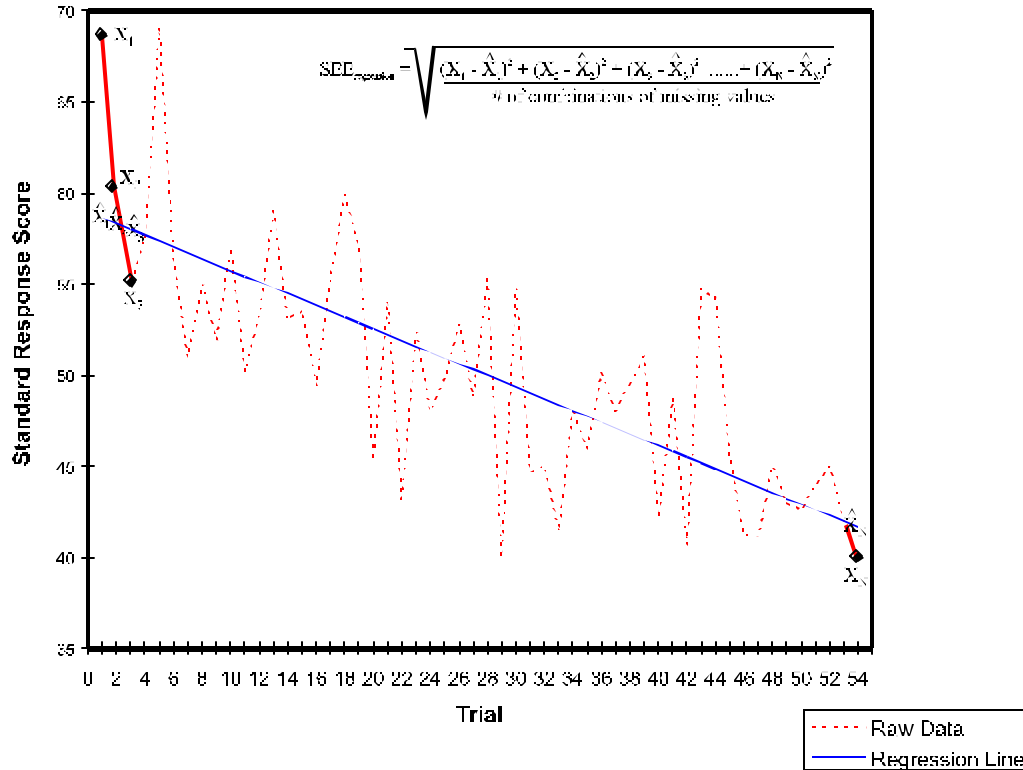
Calculation of Residual Errors for Regression and Cell Mean Methods



To compare the accuracy of these two estimation methods, a measure of the deviation of the each method's estimate from the true score was needed. We employed the standard error of estimate (SEE) to quantify the accuracy of estimation of these two methods.

In order to calculate the SEE, a residual score for each estimate was determined. A residual is the difference between the true score and the estimated score. In the figure above, the residual for the Regression estimate of X_i is the value of this estimate, X_i (depicted by the pink box), subtracted from the value of the true missing score. Similarly, the residual for the Cell Mean estimate is the value of this estimated score, C_{X_i} (depicted by the yellow box), subtracted from the true missing score.

Calculating the Standard Error of Estimate from Monte Carlo Analysis



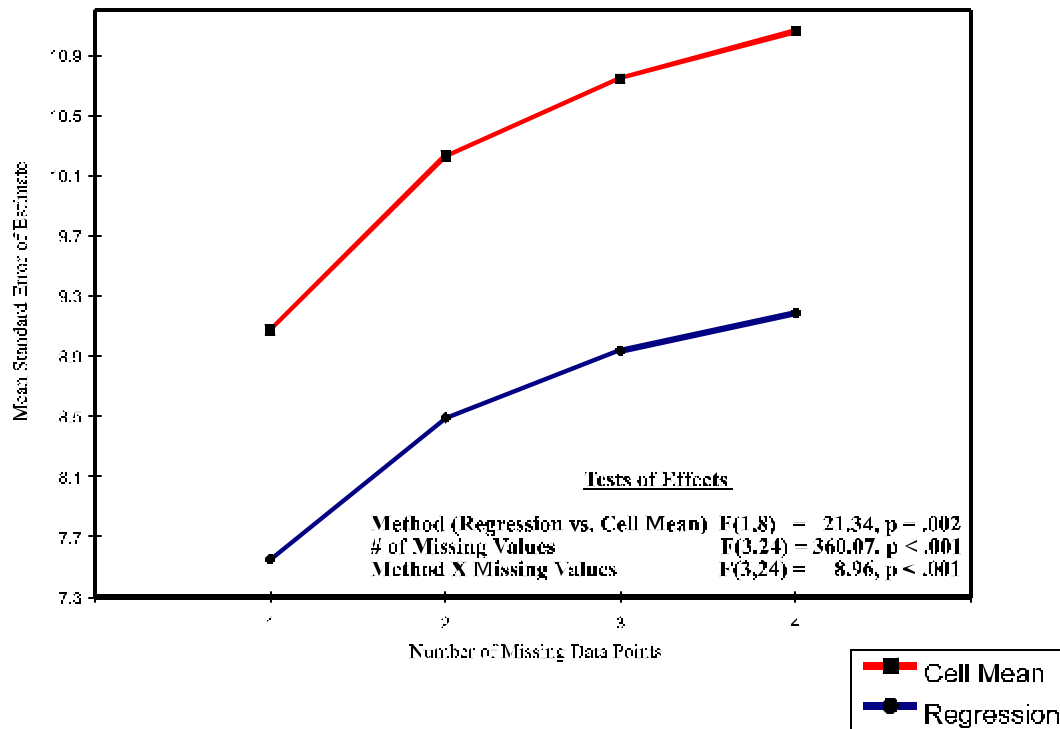
The standard error of estimate (SEE) is, in essence, the average of the residual scores. However, (as in the case of the standard deviation of scores around a mean) the average is computed for the squared residuals, and the square root is taken of the resultant.

A Monte Carlo analysis was performed to obtain the average SEE for each method across all possible combinations of missing data ranging from 1 to 4 missing scores.

In the above figure, the calculation of the SEE for a single subject's data across 54 trials for the Regression estimate for one missing data point is depicted. There are 54 possible ways that one score can be missing for this subject (i.e., trial 1 can be missing, or trial 2 can be missing... or trial 54 can be missing). Therefore, the regression SEE for one missing score for this subject is calculated by averaging the squared residuals for these 54 combinations, and then taking the square root of this average. The calculation of the SEE for the Cell Mean method is computed in the same way, except that the cell mean estimate is substituted in the calculation of the residual score.

The procedure is the same for 2, 3 and 4 missing data points, except that the number of possible combinations of missing scores increases factorially. There are 1431 possible ways that two scores can be missing from a data set of 54 trials (i.e., trials 1 & 2, or trials 1 & 3, or trials 1& 4,... or trials 53 & 54). In this case the SEE is again the square root of the average of the squared residuals across these 1431 combinations. In the cases of 3 and 4 missing data points, there are 24,804 and 316,251 possible combinations of missing values, respectively.

A Test of the Effects of Estimation Method and # of Missing Values on Standard Error of Estimate

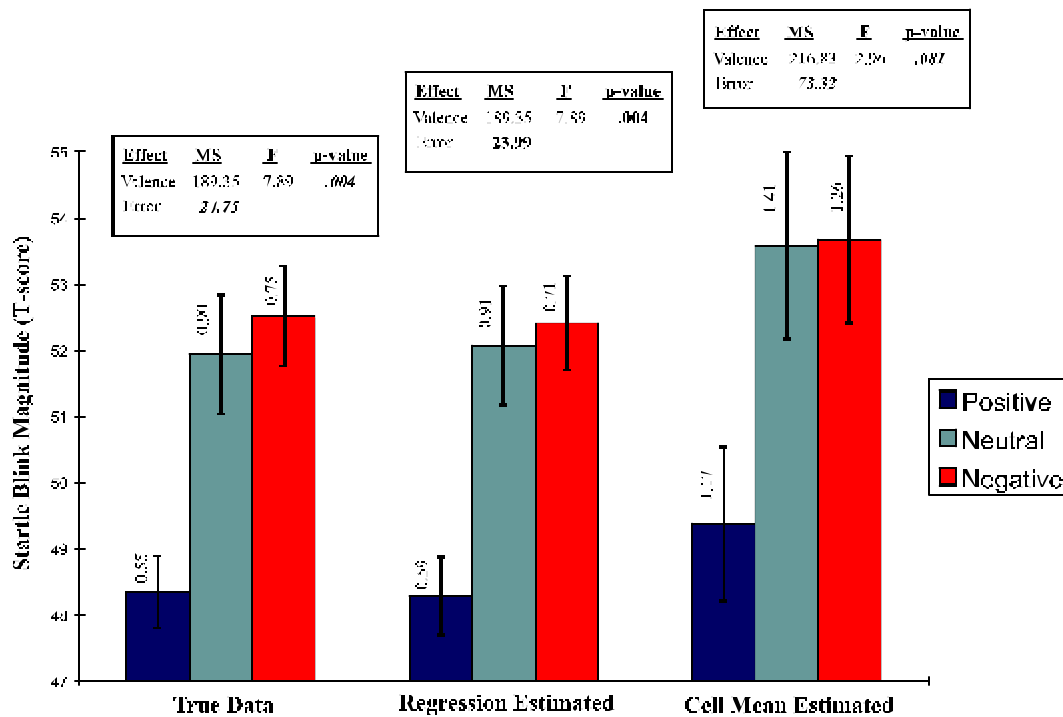


To compare the accuracy of these two estimation methods in predicting missing scores, the two methods were evaluated using a real data set. The data set came from a study investigating the effect on startle blink magnitude of slide valence (positive, neutral or negative) across six distinct probe times. Data from nine subjects with no missing values were utilized.

The Monte Carlo analysis described earlier was used to determine the SEE for each method across 1 to 4 missing data points (i.e., a score was set to missing, and each method's estimation of that score was compared to the true value of that score).

The comparison of the two estimation methods on the mean SEE as a function of number of missing scores for the 9 subjects is depicted above. The data were analyzed using a repeated measures ANOVA to determine the effect on SEE of estimation method and varying numbers of missing values. A significant main effect for Estimation method was found, with the Regression method yielding a lower SEE across the full range of missing scores, $F(1,8) = 21.34, p = .002$. Additionally, a significant main effect for number of missing scores was found, with SEE increasing with the number of missing values, $F(3,24) = 360.07, p < .001$. Finally, there was a significant interaction between Estimation method and number of missing values indicating that the improvement in accuracy of estimation for Regression as compared to Cell Mean increases with increases in the number of missing scores, $F(3,24) = 8.96, p < .001$.

A Case Example: Valence-Modulated Startle



To demonstrate concretely the potential impact of mis-estimation of missing values on subsequent data analysis, a specific example with 4 missing values per subject was generated for the valence/timing data set. Missing values were chosen to maximize the differential accuracy of the two methods.

The above figure depicts both a graphical and statistical analysis of the valence effect for the true data set (i.e., the data set with no missing values), and for the data sets with 4 missing values estimated by either the Regression or Cell Mean methods.

Statistical analysis of the true data set reveals a significant main effect for valence, $F(2,16) = 7.42$, $p = .005$. A comparable significant main effect for valence was found in the Regression estimated data set, $F(2,16) = 7.89$, $p = .004$. However, no significant effect of valence was found in the Cell Mean estimated data set, $F(2,16) = 2.96$, $p = .081$. Further examination of this discrepancy in test results of the valence effect reveals that the disparity is due to an inflated error term in the Cell Mean estimated data set. The Mean Square Error for the true and Regression estimated data sets are similar (24.75 and 23.99, respectively). However the Mean Square Error for the Cell Mean estimated data set is notably larger (73.32). This increase in error variability is a direct result of the poor estimation of missing values by the Cell Mean method.

The effect of reduced accuracy of estimation by the Cell Mean method is also apparent in the bar graphs. The means of the three levels of valence for the Regression estimated data set are qualitatively more similar to the true data set than are the means for the Cell Mean estimated data set. More importantly, however, the increase in error variability for the Cell Mean method is translated into an increase in the standard error of the means.

Conclusions

1. When estimating missing values on a dependent measure that exhibits systematic habituation, the use of a regression-based prediction incorporating trial as a predictor will yield a more accurate estimate of missing scores than simply using the cell mean.
2. The accuracy of both the Regression and the Cell Mean estimation methods decreases as the number of missing scores increases.
3. Although the accuracy of both methods degrades with increases in missing values, the reduction in accuracy is more serious for the Cell Mean method than for the Regression method.
4. The end result of increasing the accuracy of estimation of missing scores will be a reduction in the error terms of subsequent statistical analyses and more sensitive tests of experimental effects.

Future Directions

1. The utility of including additional predictors (i.e., variables other than trial, and their interactions) in the regression procedure will be investigated.
2. The effect on accuracy of estimation of incorporating between subject factors into the regression procedure will be evaluated.
3. A windows-based software application will be developed to aid in the application of the regression based estimation method to actual data sets.