# The Diffscores Macro Version 3.0

Dr Gregory John Lee

11th July 2013

## 1. BASIC DIFFSCORES MACRO V3.0 INFORMATION

Written by: Dr. Gregory John Lee, University of the Witwatersrand, Johannesburg
gregory.john.lee@gmail.com / gregory.lee@wits.ac.za
Dates: Original creation 20th March 2009, latest revision 11th July 2013.
SAS version: 9.3
Sub-macros: The Jackboot macro from SAS Support (Knowledge Base samples #24982, currently at http://support.sas.com/kb/24/982.html) is embedded in the Diffscores macro.
Disclaimer: The user employs this macro entirely at his/her own risk. Dr. Gregory John Lee and The SAS Institute take no responsibility whatsoever for its use, effects, changes made upon it or any conclusions or uses made based upon output from this macro or variants thereof

## 2. PURPOSE

The Difference Score (Diffscore) macro creates output related to analysing differences between 2 related variables as core regression predictors of a dependent variable, such as actual versus expected level of a variable as predictors of an outcome. Another example is the classic Person-Organisation Fit literature, in which the difference between the individual and organisation on various measures is of interest as a predictor of work-related outcomes.

The output helps the user choose a best model for assessment of the differences, assess the important shapes / coefficients of this model, and graph the relationships.

## 3. STATISTICAL BACKGROUND

This macro creates output based on Jeffrey R. Edwards' development of theory behind the analysis of differences between two variables as independent variables. The next section briefly describes the core concepts and models of this development.

### 3.1 Core Difference Analysis Models

Let C1 and C2 be the predictors being compared, Z be the dependent variable, and let w=1 if C1<C2 and w=0 if C1>=C2. For instance, C1 might be expected pay, C2 actual pay, and Z satisfaction. The core question is how differences between expected and actual pay affect satisfaction. There may also be a string of control/covariate variables.

Table 1 shows various models assessing differences between C1 and C2 as predictors of Z:

**Table 1: Models analyzing difference scores**

| Model | Equation |
|---|---|
| 1) Base model for comparison | $Z = controls$ |
| 2) Constrained algebraic differences | $Z = Controls + (C1-C2)$ |
| 3) Unconstrained algebraic differences (i.e. C1 and C2 separated) | $Z = Controls + C1 + C2$ |
| 4) Higher order algebraic differences (i.e. curvilinear effects, same as model 9) | $Z = Controls + C1 + C2 + C1^2 + C1xC2 + C2^2$ |
| 5) Constrained absolute differences | $Z = Controls + abs(C1-C2)$ |
| 6) Unconstrained absolute differences | $Z = Controls + C1 + C2 + w + wC1 + wC2$ |
| 7) Higher order absolute differences | $Z = Controls + C1 + C2 + w + wC1 + wC2 + C1Squared\ C1xC2\ C2Squared\ wxC1Squared\ wxC1xC2\ wxC2Squared$ |
| 8) Constrained squared differences | $Z = Controls + (C1-C2)^2$ |
| 9) Unconstrained squared differences i.e. curvilinear effects (same model as 3) | $Z = Controls + C1 + C2 + C1^2 + C1C2 + C2^2$ |
| 10) Higher order squared differences (i.e. curvilinear + cubic effects) | $Z = Controls + C1 + C2 + C1^2 + C1C2 + C2^2 + C1^3 + C1^2*C2 + C1xC2^2 + C2^3$ |

The key tasks in the analysis are to identify which of these 10 models seems to fit best and then to assess key parameters and shapes of the chosen model, including regression coefficients, graphical representation of the relationship (especially important when the relationship is unconstrained and therefore forms a 3D response surface), and shapes along or on key parts of the relationship (e.g. what the relationship is long the line of congruence, i.e. that line where C1 = C2).

### 3.2 Statistical References

Some references include:

Edwards, J. R., Cable, D. M., Williamson, I. O., Lambert, L. S., & Shipp, A. J. (2006). The phenomenology of fit: Linking the person and environment to the subjective experience of person-environment fit. *Journal of Applied Psychology*, 91, 802-827.

Edwards, J. R. (2002). *Alternatives to difference scores: Polynomial regression analysis and response surface methodology*. In F. Drasgow & N. W. Schmitt (Eds.), *Advances in measurement and data analysis* (pp. 350-400). San Francisco: Jossey-Bass.

Edwards, J. R. (2001). Ten difference score myths. *Organizational Research Methods*, 4, 264-286.

Edwards, J. R. (1994a). Regression analysis as an alternative to difference scores. *Journal of Management*, 20, 683-689.

Edwards, J. R. (1994b). The study of congruence in organizational behavior research: Critique and a proposed alternative. *Organizational Behavior and Human Decision Processes*, 58, 51-100 (erratum, 58, 323-325).

Edwards, J. R., & Parry, M. E. (1993). On the use of polynomial regression equations as an alternative to difference scores in organizational research. *Academy of Management Journal*, 36, 1577-1613.

This macro is based on the theory and arrangement in Edwards (2002), the user is strongly recommended to read that or other expositions before using this macro.

## 4. MACRO OUTPUT

The macro produces the following output. See Section 7 below for suggestions on how to analyze this output.

### 4.1 Diagnostic Outputs

The macro first outputs the following diagnostic outputs:

1) "*Diagnostic regression*": a regression analysis including all specified control, and independent variables, that includes all major PROC REG regression output (outliers, residual plots, multicollinearity, etc.);
2) "*Sorted influence scores*": A separate printout of data and influence diagnostics (CooksD and Hat) sorted descending by CooksD.

The researcher obviously is encouraged to investigate this output first and make any necessary alterations. If very influential observations exist, consider deleting or weighting these in the prior data step.

### 4.2 Model Comparisons

Second, the macro outputs the following major sections that can be used to compare models:

3) "*Main Regressions*": The PROC REG outputs of each of the major regressions. As explained in Section 7, the user will often skip reading this to start;
4) "*All Coefficients*": A table stacking all slope parameters (unstandardised & standardised and p-values) of each of the above models. The user will typically skip this initially, see Section 7;

5) "*Model Comparisons*": As per Edwards (2002) this important table reports $R^2$, F-Statistics, F-Statistic p-values and information criteria for successive comparisons of all models described in Section 3.1. The tables gives:
   a) R2 and ANOVA F (& p-value) of each of the 10 models
   b) 'R2 Change': comparison of each higher order model to the lower order model (e.g. the unconstrained algebraic compared to the constrained). The constrained models do not have this.
   c) 'FCompare' & 'pCompare': F & p stats for $R^2$ comparison with prior model in set. These equate to the Fh column in Edwards (2002).
   d) 'PC/AIC/BiC/SBC': information criteria for each model. Edwards does not have information criteria but these are most useful for choosing models as they have good parsimony adjustments. The lower the IC the better. The lowest IC in each column has a star.

   Table 2 shows how the Diffscores macro presents these model comparisons.

**Table 2: Sample of Model Comparisons Table from Diffscores Macro**

*Model comparisons Competitive_Turnover Market_Inducement*

| MODELS | DF1 | DF2 | FTEST | PVALUE | R2 | R2CHANGE | FCOMPARE | PCOMPARE | PC | AIC | BIC | SBC |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1. Base Model | 8.00 | 138.00 | 2.25 | 0.03 | 0.12 | | | | 1.00 | -375.51 | -372.35 | -348.60 |
| 2. Constrained algebraic diffs | 9.00 | 137.00 | 2.29 | 0.02 | 0.13 | 0.02 | 2.43 | 0.12 | 1.00 | -376.09 | -372.64 | -346.19 |
| 3. Unconstrained algebraic diffs | 10.00 | 136.00 | 3.70 | 0.00 | 0.21 | 0.08 | 14.36 | 0.00 | 0.91 * | -388.85* | -385.08* | -355.96* |
| 4. Higher order algebraic | 13.00 | 133.00 | 2.99 | 0.00 | 0.23 | 0.01 | 0.71 | 0.55 | 0.94 | -385.20 | -380.27 | -343.33 |
| 5. Constrained absolute diffs | 9.00 | 137.00 | 1.99 | 0.05 | 0.12 | | | | 1.01 | -373.53 | -370.08 | -343.63 |
| 6. Unconstrained absolute diffs | 13.00 | 133.00 | 3.28 | 0.00 | 0.24 | 0.13 | 5.59 | 0.00 | 0.92 | -388.39 | -383.46 | -346.52 |
| 7. Higher order absolute diffs | 19.00 | 127.00 | 2.42 | 0.00 | 0.27 | 0.02 | 0.66 | 0.68 | 0.97 | -380.92 | -372.67 | -321.11 |
| 8. Constrained squared diffs | 9.00 | 137.00 | 2.01 | 0.04 | 0.12 | | | | 1.01 | -373.70 | -370.26 | -343.80 |
| 9. Unconstrained squared diffs | 13.00 | 133.00 | 2.99 | 0.00 | 0.23 | 0.11 | 4.71 | 0.00 | 0.94 | -385.20 | -380.27 | -343.33 |
| 10. Higher order squared | 17.00 | 129.00 | 2.49 | 0.00 | 0.25 | 0.02 | 0.89 | 0.47 | 0.96 | -381.22 | -374.23 | -327.39 |

Again, see Section 7 for suggestions on the model choice process.

### 4.3 Output for Chosen Model

The macro has a "FocusModel=" option (see keywords in Section 5 below) that allows you to choose one model to analyze in depth. If the user chooses a model using this macro keyword, the macro outputs the following focusing on the chosen model:

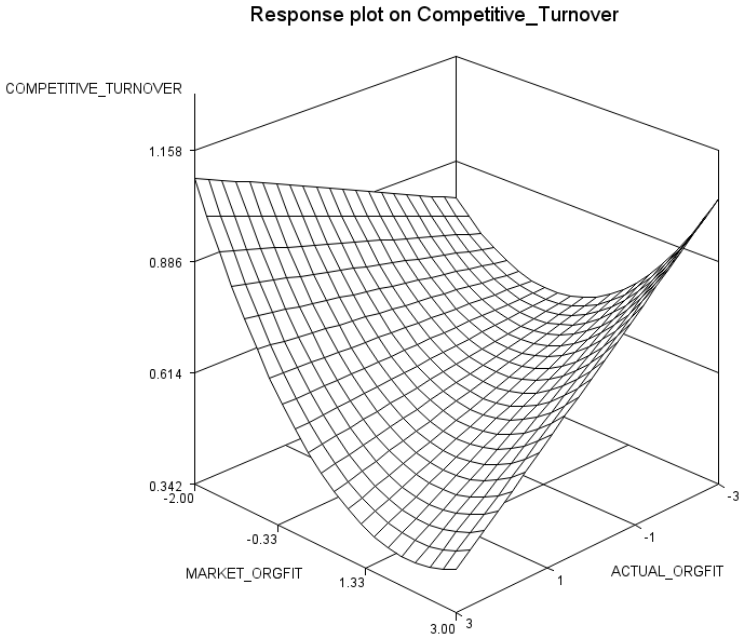
6) "*Specific coefficients*": specific regression slope coefficients and p-values of the chosen model
7) "*Locations of NB points*" (does not always show): If the user uses the FocusModel = option to focus on the unconstrained squared differences model specifically, this table outputs the location of important points, namely the stationary point and first and second principal axes (PAs). Read the material by Edwards (2002) for more on this, in the output X0 = value of X at the stationary point, Y0 = value of Y at the stationary point, P11 = slope of 1st principal axis (PA), P10 = intercept of 1st PA, P21 = slope of 2nd PA, P20 = intercept of 2nd PA6)
8) "*Shapes of NB lines*" (does not always show): If the user uses the "FocusModel=" option to focus on either the unconstrained squared differences model (Model 4 or 9) or unconstrained algebraic models (Model 3) specifically, then this table outputs:
   a) The linear slope estimate for important lines, namely the linear slopes along the lines of perfect congruence X=Y, and perfect incongruence Y = -X, and (in the case of the unconstrained squared model) the principal axes.
   b) For the unconstrained squared model the coefficient of cuvilinearity in these lines is given (coefficients ax12, ax22, ax32 and ax42).

   The bootstrap option can be invoked to get confidence intervals for these shapes.
9) "Initial Bootstrap" outputs (If 'Bootstrap =' option is Y or blank): Initial bootstrap results with number of resamples and alphas designated by the researcher, typically these are ignored except to check effect of bootstrap

10) "Percentile bootstrap" output (If 'Bootstrap =' option is Y or blank): The percentile bootstrap confidence intervals
11) "BcA bootsrap (If BCA = Y is requested): The Bca bootstrap confidence intervals for comparison
12) "Graph" sections (If Graph = Y is requested with a FocusModel = option ): SAS will graph the shape chosen by the user in the FocusModel = option.
    a) For higher order models the graph will be a response surface (which the user may rotate using the Rotate = option to get the best view)
    b) For the simple models (constrained algebraic, absolute or squared models) the macro plots a simple 2D graph of the relationship.

    Figure 1 shows a sample of the graph printout.
    Ignore the first graph section which is just the initial scatter extrapolation used to graph the final surface.

**Figure 1: Sample of Graph Output from Diffscores macro**




## 5. INSTRUCTIONS FOR MACRO USE

The User should undertake the following steps – also see Section 7 for suggestions for a user process.

### 5.1 Data Preparation

The usual data considerations for polynomial regression apply. The user may usually wish to center his/her independent variables before using them, undertake missing value analysis, and the like.

### 5.2 Save the Macro to File and Include in Program Editor

Having saved the macro file on a given location, include the macro in your program, usually by using the command *%inc'<file pathname here>';* for instance:

*%inc'C:|Users|GregLee|Documents|Statistics|My SAS macros|DiffScores.sas';*

### 5.3 Invoke the Diffscores Macro and Keywords

Invoke the macro by typing *%Difference(<insert keywords here>).* The following sections describe the required and optional keywords (see Section for sample code):

a) *Required Keywords for Diffscores Macro*

1) *Data=* (input dataset name)
2) *,Comparison1=* (give the variable name for the first comparison independent variable. I suggest whichever of the two variables is the natural benchmark, e.g. in a comparison between "Actual" and "Optimal" Person-Organisation Fit, pick the "Optimal" here.)
3) *,Comparison2=* (give any name for the second comparison independent variable that is being compared with the first.)
4) *,Depvar=* (The name of the dependent variable, it MUST correspond with the actual variable name in the dataset)

b) *Optional Keywords for Diffscores Macro*

5) *,Controls=* (Any other regression variables you wish to include as covariates of the main comparison variables. These must be named using the variable names in the dataset)
6) *,IDvar =* (Any variable that is used as a row identity variable such as an employee or survey number). I highly recommend using an index variable as it facilitates identification of outliers in the second output;
7) *,RegModelOptions=* (By default the macro asks for standardised coefficients, if the researcher wants extra MODEL options such as White's covariance adjustment, these should be listed here using the correct PROC REG keywords)
8) *,FocusModel =* This important keyword identifies for which model the macro should print specific coefficients, bootstrap (if chosen), and graph (if chosen). The user should put a single number from 1-10 corresponding with the model numbers in the Table in Section 3.1 on p. 2 above (or in the Model Comparisons output table). For instance, *FocusModel = 9* will produce output focusing on the uncontrained squared differences model.
   Note the FocusModel option cannot take multiple options, either leave blank or give one number. If left blank, the macro does not output any focused tables and terminates at model comparisons.
9) *,Bootstrap=* (Options are Y or N, if a FocusModel option is chosen then default Bootstrap = option is Y which causes bootstrap results to run with default 10000 iterations and 95% percentile confidence intervals. Number of resamples, alpha level and whether to also invoke Bca confidence levels can be altered with the options below)
10) *,Resamples=* (Number of bootstrap resamples to draw, 10000 is the default)
11) *,Bootstrapalpha=* (nominate alpha level for bootstrap confidence levels etc. Default is .05 for 5%)
12) *,Bca=* (options are Y or N, Y will ask SAS to include Bca confidence intervals for comparison. These adjust for both bias and acceleration, but the researcher is warned that they are computationally heavy, so for large samples this option may be slow, Default is N)
13) *,Graph =* (Default is Y if a FocusModel= option is chosen. Graph = Y will create a response surface using PROC G3DGRID and PROC G3D for all options except for the constrained differences in which case the macro outputs a 2D graph using GPLOT. The data chosen is a random combination of datapoints within the original range of maxima and minima of the predictor variables, with the corner extremes always included to ensure coverage.
14) *,Rotate=* The user can rotate a response surface graph by stipulating and angle here, e.g. Rotate = 225 obviously rotates to 225 degrees.

## 6. SAMPLE CODE

```
<datastep here if necessary>
%inc'<location of DIFFSCORES macro>';
%Difference(
Data=Destinations
,Comparison1=Market_Pay
,Comparison2=Actual_Pay
,DepVar=Competitive_Turnover
,Controls=services finance topman POSten age edu white male
,IDvar=Resp
,RegModelOptions=
```

,FocusModel = 3
,bootstrap = y
,BCA = y
,resamples = 10000
,bootstrapalpha = .05
,Graph = Y
,Rotate = 225)

This will produce bootstrapped output with 10000 resamples, .05 percentile and BCA CIs, and a graph.

## 7. SUGGESTIONS FOR USER PROCESS

I suggest that the user does the following in the process of using the macro for analysis:

1. Run the macro without a FocusModel option.
2. Analyze the diagnostic regression and influence scores. Decide whether the basic regression is relatively sound in terms of traditional regression assumptions. If necessary, deal with influential points and other changes (e.g. you may wish to transform variables).
3. Jump to the Model Comparisons table. Decide which model, if any, appears to fit well / best. Compare models using R-square comparisons and information criteria.
4. If you decide on a seeingly best model, reverse to the Main Regressions section and assess this specific regression, if necessary adding model options in the "RegModelOptions=" keyword section. Ensure you are happy with the chosen model. Important: check that there seems to be some reasonable size / significance to the key parameters of the chosen model. This stipulation is necessary because sometimes a model will seemingly be superior due to a significantly higher r-square in the Model Comparisons table, but none of the extra, key parameters have any real magnitude and are non-significant. (For instance, you may choose the unconstrained squared model)but none of the higher order C1squared, C1xC2 or C2squared parameters are substantial / significant.) Note that the information criteria usually avoid this.
5. Re-run the macro with the "FocusModel=" keyword set to the chosen model. I now suggest choosing Y for the "Bootstrap=" and "Graph=" options. At this stage a full analysis of the chosen model is possible.
6. Write up the analysis as per Edwards (2002) or other similar analyses. Remember that if you choose the unconstrained algebraic or squared models, have the extra table(s) showing the shapes along the important lines (e.g. slope coefficients for lines of congruence and incongruence), and use the bootstrap to show relative significance of those shapes. (I run the bootstrap at 3 levels of significance for comparison).