Using SAS® to Test, Probe and Display Interaction Effects in Linear Models
Timothy B. Gravelle, PriceMetrix Inc., Toronto, Ontario, Canada

ABSTRACT

Interaction effects (or moderated effects) in regression models capture how the effect of an independent variable on the dependent variable varies as a function of a third variable. In general terms, interaction effects are interpreted as the effect of $X_1$ on $Y$ at different values of $X_2$. Such interactive (or moderated) relationships pervade statistical research in a wide variety of disciplines. Still, their specification and communicating their substantive interpretation remains a challenge.

This paper demonstrates how SAS® software can be used to specify, probe and display interaction effects in linear regression models. It covers a number of techniques, including different graphical displays for interaction effects, and techniques for identifying the regions of statistical significance for an independent variable.

INTRODUCTION

The application of modern regression methods in numerous fields of study – from econometrics to political behavior and public health – often entails the use of interaction effects (also called moderator effects). The term “interaction effect” describes a situation in which the effect of an independent variable on the dependent is conditional upon the value of another variable, usually termed a moderator variable (Aiken and West 1991; Hayes et al. 2012; Jaccard and Turrisi 2003; Jaccard and Dodge 2004). In other words, “interactive relationships imply that the impact of $X_1$ on $Y$ varies depending on the level of $X_2$” (Braumoeller 2004: 809).

Despite their seeming ubiquity, more than one author has noted that such effects are often misinterpreted in applied research (Braumoeller 2004; Edwards 2009; Hayes et al. 2011). Part of the challenge in interpreting interaction effects is that a set of regression coefficients often fails to communicate plainly the substantive meaning of an interaction effect. Several solutions to this challenge have been proposed, including graphical displays of interactive effects (Fox 1987) and plots of estimated coefficients and their confidence intervals (Braumoeller 2004). These solutions are easily implemented with SAS® software.

In this paper, I cover the use of SAS/STAT® and SAS/GRAPH® to specify, test, probe and display interactive (or moderated) relationships in regression models. After presenting my research topic, I review some basic regression equations and the general theoretical framework for interaction effects. I then discuss data preparation for testing interactive effects, creating graphical displays for interaction effects, and techniques for identifying the regions of statistical significance for an independent variable. My focus is on the familiar linear model as implemented by ordinary least squares (OLS), though both the theory and application can be readily extended to generalized linear models (e.g., logit and probit models, and models for count data).

RESEARCH TOPIC: PREDICTING INVESTMENT ADVISORS’ PRODUCTIVITY

This paper is motivated by a practical research question: how can one predict investment advisors’ future productivity (measured in revenue) given data on advisors’ performance in a baseline period, personal characteristics and information on the advisors’ client base? To answer this question, I draw on data from the proprietary PriceMetrix retail wealth management database. My dataset consists of 1,010 investment advisors from multiple firms in North America with 5 to 15 years of industry experience as of the end of 2006.

My dependent variable is total revenue in thousands of dollars for the year 2011 (TOTAL_REV_11). My independent variables are a set of advisor and advisor book characteristics taken from 2006. These are:
- transactional (trading commission) revenue in thousands of dollars (TRANS_REV_06);
- mutual fund trailer revenue in thousands of dollars (TRAILER_REV_06);
- revenue for fee-based services in thousands of dollars (FEE_REV_06);
- years of experience (EXPERIENCE_YEARS);
- a dichotomous variable indicating whether an advisor practices as part of a team (TEAM);
- the count of large (or “core”) households in an advisor’s book (CORE_HH_COUNT);
- the count of small households in an advisor’s book (SMALL_HH_COUNT);
- the number of retirement savings accounts in an advisor’s book (RETIREMENT_ACCT_COUNT);
- the average number of accounts per household in an advisor’s book (ACCTS_PER_HH), serving as a measure of household share of wallet; and,
- median household assets (MEDIAN_HH_ASSETS), a measure of average household size for an advisor’s book.

All of these variables (measured at time 0) are hypothesized predictors of future productivity (measured at time 0 + 5). (Basic descriptive statistics for these data appear in Fig. 1.) Additionally, subject knowledge suggests that the effect of household size (MEDIAN_HH_ASSETS) on future revenue may depend on the breadth and depth of an advisor’s client relationships (or client share of wallet), a measure of which is accounts per household. This implies an interaction effect.

---

### Advisor Data: Basic Descriptive Statistics

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Median</th>
<th>Std Dev</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOTAL_REV_11</td>
<td>760.17</td>
<td>14.61</td>
<td>3888.97</td>
<td>630.72</td>
<td>481.83</td>
<td>1010</td>
</tr>
<tr>
<td>TRANS_REV_06</td>
<td>327.19</td>
<td>7.40</td>
<td>2184.26</td>
<td>253.86</td>
<td>273.46</td>
<td>1010</td>
</tr>
<tr>
<td>TRAILER_REV_06</td>
<td>101.43</td>
<td>0.00</td>
<td>726.07</td>
<td>77.91</td>
<td>86.99</td>
<td>1010</td>
</tr>
<tr>
<td>FEE_REV_06</td>
<td>199.22</td>
<td>0.00</td>
<td>2443.41</td>
<td>101.16</td>
<td>257.27</td>
<td>1010</td>
</tr>
<tr>
<td>EXPERIENCE_YEARS</td>
<td>11.96</td>
<td>5.05</td>
<td>20.00</td>
<td>11.81</td>
<td>4.27</td>
<td>1010</td>
</tr>
<tr>
<td>TEAM</td>
<td>0.06</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
<td>0.24</td>
<td>1010</td>
</tr>
<tr>
<td>CORE_HH_COUNT</td>
<td>74.17</td>
<td>1.00</td>
<td>383.00</td>
<td>65.00</td>
<td>44.75</td>
<td>1010</td>
</tr>
<tr>
<td>SMALL_HH_COUNT</td>
<td>173.33</td>
<td>4.00</td>
<td>1269.00</td>
<td>142.00</td>
<td>134.40</td>
<td>1010</td>
</tr>
<tr>
<td>RETIREMENT_ACCT_COUNT</td>
<td>127.65</td>
<td>0.00</td>
<td>1141.00</td>
<td>112.00</td>
<td>102.08</td>
<td>1010</td>
</tr>
<tr>
<td>ACCTS_PER_HH</td>
<td>2.08</td>
<td>1.07</td>
<td>4.01</td>
<td>2.08</td>
<td>0.40</td>
<td>1010</td>
</tr>
<tr>
<td>MEDIAN_HH_ASSETS</td>
<td>150.71</td>
<td>0.10</td>
<td>715.06</td>
<td>139.55</td>
<td>87.65</td>
<td>1010</td>
</tr>
</tbody>
</table>

---

**THEORIZING AND SPECIFYING INTERACTION EFFECTS**

Conceptualizing an interaction effect is often aided by thinking in terms of a system comprising three variables: the dependent variable, the “focal” independent variable and the moderator variable. The effect of the focal independent variable on the dependent variable is said to be moderated by the moderator variable (Jaccard and Turrisi 2003; Jaccard and Dodge 2004). Applying this framework to the case of investment advisors’ productivity, one can say that the relationship between productivity (dependent variable) and median household assets (independent variable) is contingent upon (or moderated by) the average number of accounts per household (moderator variable). Thus, median household size and number of accounts per household interact.

It is important to distinguish interaction effects (also termed moderated effects) from mediated effects (also termed indirect effects). The classic statement on moderation versus mediation comes from Baron and Kenny, in which they write that a moderator variable is one “that affects the direction and/or strength of the relation between an independent or predictor variable and a dependent or criterion variable” (1986:...
1174). Conversely, a variable “may be said to function as a mediator to the extent that it accounts for the relation between the predictor and the criterion.” This is to say that when one controls for the effect of the mediator variable on the dependent variable, the effect of the (hypothesized) independent variable is zero (Baron and Kenny 1986: 1176). These are two distinct patterns of effects at both the level of theory and their statistical specification and testing (see Figs. 2a and 2b). I am concerned only with moderated (or interactive) relationships in this paper.

The specification of interaction effects in linear models is typically done using product terms. Assuming a linear model with two independent variables and an interaction term, this model may be written as:

\[ \hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2, \]

where \( \hat{Y} \) is the predicted value (conditional mean) of the dependent variable, \( \beta_0 \) is the intercept term, \( \beta_1 \) and \( \beta_2 \) represent the regression coefficients for the independent variables in the model, \( \beta_{12} \) is the coefficient for the interaction term, and \( X_1 \) and \( X_2 \) represent the values taken by the independent variables. Typically, \( \beta_1 \) and \( \beta_2 \) are referred to as the lower-order terms comprising an interaction effect and \( \beta_{12} \) is referred to as the higher-order term. This example is also a case of two-way interaction effect (it involves only two variables). Three-way (and higher) interaction effects are possible, though they are more difficult to interpret and are rarer in published research.

Specifying such a model in PROC REG is straightforward. At its most basic, it may be specified as shown in Fig. 3.

At this point, some observations about the above model are necessary. First, when hypothesizing an interaction effect, the parameters comprising the lower-order terms of the interaction (\( \beta_1 \) and \( \beta_2 \) in this example) must still be present in the model. Otherwise, the model is not “hierarchically well-formulated” and is not of general application (Jaccard and Dodge 2004: 240).

Second, even when lower-order terms are included in the model, they are not of primary interest. Moreover, as numerous authors have stressed, they are not interpretable as the “main effects” of \( \beta_1 \) and \( \beta_2 \) or the “effects of \( \beta_1 \) and \( \beta_2 \) in general” or the “average effects” of \( \beta_1 \) and \( \beta_2 \); models including interaction terms do not allow for such an interpretation. Strictly speaking, the lower-order terms indicate the effect of \( X_1 \) on \( \hat{Y} \) when \( X_2 \) is 0, and similarly the effect of \( X_2 \) on \( \hat{Y} \) when \( X_1 \) is 0 (Braumoeller 2004: 809; Edwards 2008: 146–148; Fox 2008: 133–134; Friedrich 1982: 821). Depending on one’s data, the value zero may be out of the range of possible values for \( X_1 \) and \( X_2 \), and so such results are of little practical interest. Still, the value zero can be made meaningful by additive transformations of the independent variables – a point developed in the next section (cf. Allison 1977: 145–148; Edwards 2008: 145–146).

**DATA PREPARATION: “PROPER CARE AND FEEDING” OF INTERACTIONS**

When specifying and testing linear models with interaction effects, all of the standard assumptions underpinning linear regression – including linearity in the predictors, normality, constant error variance (homoscedasticity), independence of the errors and absence of high collinearity among the predictors – continue to apply. A detailed discussion of these topics is beyond the scope of this paper (see Fox 2008).

One data preparation technique that is especially relevant when modeling interaction effects and worth discussing, though, is **mean-centering**. This simply involves subtracting the arithmetic mean from the original scores, resulting in new scores with a mean of zero. Jaccard and Turrisi recommend this strategy
as a way to “force the coefficients to reflect parameters that are of theoretical interest” (2003: 15). Often, the value zero does not occur within the range of one’s data, or it is an extreme value. In such cases, OLS regression coefficients that capture an independent variable’s effect with all other independent variables set at zero hold little substantive interest. It is better to transform the data so that the value zero is in fact meaningful – for example, by taking on the interpretation of a variable’s pre-transformation mean value. Mean-centering accomplishes this. (Median-centering is also an option, as are other values having a theoretical basis as chosen by the analyst.) In addition to theoretical considerations, Dalal and Zickal (2012) and Robinson and Schumacker (2009) also recommend mean-centering the independent variables to reduce collinearity between the lower-order and higher-order variables comprising the interaction effect, though Edwards (2008) points out that mean-centering otherwise has no impact on the estimation or interpretation of the regression model itself.

Mean-centering is easily accomplished in SAS® using PROC STDIZE and the METHOD=MEAN option (the default method is to create standardized scores). A PROC STDIZE run that creates a new dataset with mean-centered independent variables (with the exception of the dichotomous variable TEAM) is shown in Fig. 4.

Once the independent variables are mean centered, the product term for ACCTS_PER_HH and MEDIAN_HH_ASSETS can be created in a simple data step as shown in Fig. 5. Creating product terms in a data step is necessary when modeling using PROC REG. Unlike PROC GLM, interaction terms cannot be entered directly into PROC REG. Another PROC MEANS run on the new dataset created by PROC STDIZE confirms that the continuous independent variables now have means of zero, which again represent the mean values from the original data.
LET'S INTERACT!

With the data prepared for analysis – continuous independent variables are mean-centered and the product term created to test the interaction between MEDIAN_HH_ASSETS and ACCTS_PER_HH – the regression models can be specified and tested using PROC REG. Typically, the model is first fit without the interaction term. The model is then refit with the interaction term included. This two-step process is more important when the interaction comprises more than one degree of freedom and omnibus tests of model fit are required to assess the statistical significance of the interaction, but it is a good practice in all instances (cf. Edwards 2008: 150–151).

The PROC REG run to accomplish this appears in Fig. 7. It is important to note the use of the OUTEST option in the PROC REG statement which outputs the model coefficients to a new dataset, parmest. This will prove useful in creating effects plots later on. One can additionally request diagnostic plots, adjusted $R^2$, confidence intervals for the regression coefficients, standardized coefficients and variance inflation factors to assess collinearity (among other options) as one sees fit. Partial output from the parmest dataset appears in Fig. 8, and partial output from PROC REG is reproduced in Fig. 9.

Setting aside a detailed discussion of each regression coefficient, I concentrate here on overall model performance and the effects of the two variables involved in the interaction effect. Examining the output for the first (main effects) model, we see that the overall model is highly predictive ($R^2 = 0.812$, $p < 0.0001$). Due to the mean-centering of the continuous independent variables, the model intercept (751.27) has a meaningful interpretation: it is the expected value of TOTAL_REV_11 for an investment advisor who does not practice in a team setting (TEAM=0) and whose scores on all of the independent variables are at their respective means. One would therefore expect such an “average” advisor to produce $751,270 in revenue for 2011.

Also, one sees that the effect of ACCTS_PER_HH is significant ($p = 0.016$). The coefficient for ACCTS_PER_HH indicates that, controlling for other factors, an investment advisor with 3.08 accounts per household (one account per household more than the sample average of 2.08) is expected to produce $53,813 more revenue than an advisor with an average accounts-to-household ratio. The coefficient for MEDIAN_HH_ASSETS (-0.114), however, is insignificant ($p = 0.315$). This result indicates that there is insufficient evidence to conclude that advisors whose clients are typically larger produce more revenue than advisors whose clients are typically smaller. In light of subject knowledge, which predicts a positive and significant effect, this result is counter-intuitive. It may be that the assumption of a constant effect of MEDIAN_HH_ASSETS for all values of ACCTS_PER_HH does not hold. Rather, the effect of MEDIAN_HH_ASSETS may be contingent upon the value of ACCTS_PER_HH. This is what the interaction effect model tests.

![Fig. 7]

```plaintext
PROC REG DATA=data_2 OUTEST=parmest;
MAIN EFFECTS: MODEL TOTAL_REV_11 = TRANS_REV_06 TRAILER_REV_06 FEE_REV_06
  EXPERIENCE_YEARS TEAM CORE_HH_COUNT SMALL_HH_COUNT RETIREMENT_ACCT_COUNT
  ACCTS_PER_HH MEDIAN_HH_ASSETS
  /ADJSQ CLB STB VIF;
INTERACTION: MODEL TOTAL_REV_11 = TRANS_REV_06 TRAILER_REV_06 FEE_REV_06
  EXPERIENCE_YEARS TEAM CORE_HH_COUNT SMALL_HH_COUNT RETIREMENT_ACCT_COUNT
  ACCTS_PER_HH MEDIAN_HH_ASSETS ACCTS_PER_HH_MED_HH_ASSETS
  /ADJSQ CLB STB VIF;
INT_EFFECT: TEST ACCTS_PER_HH_MED_HH_ASSETS=0;
RUN; QUIT;
```
Fig. 8

<table>
<thead>
<tr>
<th>Obs</th>
<th><em>MODEL</em></th>
<th><em>TYPE</em></th>
<th><em>DEPVAR</em></th>
<th>Intercept</th>
<th>ACCTS_PER_HH</th>
<th>MEDIAN_HH_ASSETS</th>
<th>HH_MED_HH_ASSETS</th>
<th>ACCTS_PER_HH</th>
<th><em>RSQ</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>MAINEFFECTS</td>
<td>PARM</td>
<td>TOTAL_REV_11</td>
<td>751.270</td>
<td>53.813</td>
<td>-0.114</td>
<td>.</td>
<td>0.812</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>INTERACTION</td>
<td>PARM</td>
<td>TOTAL_REV_11</td>
<td>739.794</td>
<td>46.247</td>
<td>-0.305</td>
<td>0.562</td>
<td>0.814</td>
<td></td>
</tr>
</tbody>
</table>

Fig. 9

The REG Procedure
Model: MAINEFFECTS
Dependent Variable: TOTAL_REV_11

Number of Observations Read 1010
Number of Observations Used 1010

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>10</td>
<td>190238759</td>
<td>19023876</td>
<td>431.86</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>999</td>
<td>44007296</td>
<td>44051</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>1009</td>
<td>234246055</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE  209.88413  R-Square  0.8121
Dependent Mean  760.17482  Adj R-Sq  0.8103
Coeff Var  27.60998

Parameter Estimates

| Variable                | DF | Parameter Estimate | Standard Error | t Value | Pr > |t| |
|-------------------------|----|--------------------|----------------|---------|-------|
| Intercept               | 1  | 751.26991          | 6.85512        | 109.59  | <.0001|
| TRANS_REV_06            | 1  | 0.82323            | 0.02916        | 28.24   | <.0001|
| TRAILER_REV_06          | 1  | 0.77207            | 0.09913        | 7.79    | <.0001|
| FEE_REV_06              | 1  | 0.97834            | 0.03272        | 29.90   | <.0001|
| EXPERIENCE_YEARS        | 1  | -12.56149          | 1.62893        | -7.71   | <.0001|
| TEAM                    | 1  | 142.76120          | 29.46273       | 4.85    | <.0001|
| CORE_HH_COUNT           | 1  | 1.32491            | 0.27042        | 4.90    | <.0001|
| SMALL_HH_COUNT          | 1  | -0.15372           | 0.11017        | -1.52   | 0.1290|
| RETIREMENT_ACCT_COUNT   | 1  | 0.47258            | 0.11506        | 4.11    | <.0001|
| ACCTS_PER_HH            | 1  | 53.81291           | 22.22005       | 2.42    | 0.0156|
| MEDIAN_HH_ASSETS        | 1  | -0.11352           | 0.11302        | -1.00   | 0.3154|
The REG Procedure
Model: INTERACTION
Dependent Variable: TOTAL_REV_11

Number of Observations Read 1010
Number of Observations Used 1010

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>11</td>
<td>190669800</td>
<td>17333618</td>
<td>396.98</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>998</td>
<td>43576255</td>
<td>43664</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>1009</td>
<td>234246055</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE 208.95833
R-Square 0.8140
Dependent Mean 760.17482
Adj R-Sq 0.8119
Coeff Var 27.48819

Parameter Estimates

| Variable            | DF | Parameter Estimate | Standard Error | t Value | Pr > |t| |
|---------------------|----|--------------------|----------------|---------|------|
| Intercept           | 1  | 739.79385          | 7.74079        | 95.57   | <.0001 |
| TRANS_REV_06        | 1  | 0.80831            | 0.02941        | 27.48   | <.0001 |
| TRAILER_REV_06      | 1  | 0.76138            | 0.09875        | 7.71    | <.0001 |
| FEE_REV_06          | 1  | 0.96004            | 0.03309        | 29.01   | <.0001 |
| EXPERIENCE_YEARS    | 1  | -12.66608          | 1.62208        | -7.81   | <.0001 |
| TEAM                | 1  | 140.36661          | 29.34267       | 4.78    | <.0001 |
| CORE_HH_COUNT       | 1  | 1.64718            | 0.28810        | 5.72    | <.0001 |
| SMALL_HH_COUNT      | 1  | -0.26999           | 0.10731        | -2.52   | 0.0120 |
| RETIREMENT_ACCT_COUNT| 1 | 0.50773            | 0.11510        | 4.41    | <.0001 |
| ACCTS_PER_HH        | 1  | 46.24691           | 22.5271        | 2.08    | 0.0379 |
| MEDIAN_HH_ASSETS    | 1  | -0.30548           | 0.12804        | -2.39   | 0.0172 |
| ACCTS_PER_HH_MED_HH_ASSETS | 1 | 0.56231            | 0.17897        | 3.14    | 0.0017 |

Test INT_EFFECT Results for Dependent Variable TOTAL_REV_11

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Numerator</td>
<td>1</td>
<td>431040</td>
<td>9.87</td>
<td>0.0017</td>
</tr>
<tr>
<td>Denominator</td>
<td>998</td>
<td>43664</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Turning to the interaction model, we find that it improves on the main effects model. Two results confirm this: the coefficient of the interaction term (and its standard error) and the $F$-ratio test produced by the TEST statement. In actuality, these two tests provide equivalent results. The coefficient of 0.562 for ACCTS_PER_HH_MED_HH_ASSETS is statistically significant ($p = 0.0017$). From this we can conclude that the effect of MED_HH_ASSETS is contingent upon the value of ACCTS_PER_HH. In other words, ACCTS_PER_HH moderates the effect of MEDIAN_HH_ASSETS. It is also worth noting that given presence of the interaction effect in the model, the coefficients for MEDIAN_HH_ASSETS and ACCTS_PER_HH are not of primary interest. Strictly speaking, they are interpreted as the effect of MEDIAN_HH_ASSETS when ACCTS_PER_HH is zero – actually, its mean on account of being mean-centered – and conversely the effect of ACCTS_PER_HH when MEDIAN_HH_ASSETS is zero – again, actually its mean (cf. Edwards 2008: 148).

The $F$-ratio test produced by the TEST statement in PROC REG indicates that the interaction term is significantly different from zero. The $p$ value for this $F$-ratio is the same as that for the estimate of the ACCTS_PER_HH_MED_HH_ASSETS (0.0017). This is what we would expect, as $F$ is simply $t^2$ (9.87 = $3.14^2$) (Edwards 2008: 150). The TEST statement, then, would appear to provide redundant information. This is only the case, however, where an interaction effect comprises a single degree of freedom, as it does here. The utility of the TEST statement is most obvious when modeling an interaction effect involving a polytomous (multicategory) variable requiring multiple dummy variables (and thus multiple degrees of freedom). In such cases, the correct test of the statistical significance of the interaction is the $F$-ratio test and not the significance of the individual coefficients, as it is possible to have a statistically significant interaction as judged by the $F$-ratio test even if some (or all) of the coefficients comprising the interaction are seemingly insignificant (Jaccard and Turrisi 2003: 40).

A PLOT IS WORTH A THOUSAND WORDS (OR COEFFICIENTS)

While the results above confirm a significant interaction between MEDIAN_HH_ASSETS and ACCTS_PER_HH, its interpretation remains unclear. The interaction must be further probed to clarify its substantive meaning. Different strategies exist to depict graphically the moderated relationship between the focal independent variable and dependent variable. The first strategy involves producing an effect plot highlighting the strength and direction of the relationship between the focal independent variable and dependent variable at different levels of the moderator variable. The second complementary strategy involves producing a plot displaying the estimated coefficient for the focal independent variable (and its confidence interval) with the scores for the moderator variable centered at different values. This serves to highlight the regions of significance of the focal independent variable – that is, the range of values of the moderator variable over which the focal independent variable exerts a significant effect. I shall refer to this type of plot as a coefficient plot. I discuss each in turn below.

EFFECT PLOTS

The principle underlying effect plots (or effect displays) is to allow the variables comprising the interaction effect to range over their values while holding other variables constant at some typical value (such as their means) and then to use the regression equation to calculate the predicted value of the dependent variable. These predicted values are then graphed (Aiken and West 1991: 12–15; Fox 2008: 136–137). Applied to the case investment advisor productivity, one could calculate predicted values of TOTAL_REV_11 for (uncentered) values of MEDIAN_HH_ASSETS between 100 and 600 and (uncentered) values of ACCTS_PER_HH of 1.5 to 4.0 in intervals of 0.5 and then plot the results.

The step-by-step process to create such an effect plot is as follows, and the program for accomplishing these tasks appears in Fig. 10. First, the means for ACCTS_PER_HH and MEDIAN_HH_ASSETS are output from PROC UNIVARIATE to a new dataset named means. Their mean values are then assigned to the macro variables AVG_ACCTS_PER_HH and AVG_MEDIAN_HH_ASSETS in a DATA step. In the following DATA step, the predicted values for TOTAL_REV_11 are created in a new dataset named plot_1. Here, the regression coefficients saved in the parnest dataset are iteratively processed in nested
DO loops that generate the desired values for ACCTS_PER_HH and MEDIAN_HH_ASSETS and then multiply these values out by the coefficients for MEDIAN_HH_ASSETS, ACCTS_PER_HH and the interaction effect ACCTS_PER_HH_MEDIAN_HH_ASSETS. The last DATA step creates a new dataset named plot_2 that restructures the plot_1 dataset using MERGE and BY statements, creating a single row for each value of MEDIAN_HH_ASSETS and multiple columns for the different values of ACCTS_PER_HH. The plot_2 dataset then serves as the input to the PROC SGPLOT run in Fig. 11. The output graph from PROC SGPLOT appears in Fig. 12.

Examining this plot, it is clear that the relationship between productivity and median household size is stronger (and positive) when accounts per household is high (when ACCTS_PER_HH is in the range of 3.5 to 4.0). When ACCTS_PER_HH is in the range of 2.5 to 3.0, the relationship between productivity and median household size is relatively weak. Such shallow slopes also suggest that the relationship may not be significantly different from zero for such values of ACCTS_PER_HH, and the data should be further probed to ascertain this. Lastly, when ACCTS_PER_HH is in the range of 1.5 to 2.0, the relationship between productivity and median household size is negative.

Substantively, then, the conventional wisdom in the wealth management industry that high-asset households are good for an investment advisor’s book of business is shown to be highly contingent. Predicted advisor productivity is highest when an advisor’s book of business contains many households with significant assets (indicated by high MEDIAN_HH_ASSETS) and an advisor has deep relationships with his or her clients (indicated by high ACCTS_PER_HH). Where an advisor’s relationships are relatively shallow (indicated by low ACCTS_PER_HH), having high MEDIAN_HH_ASSETS actually leads to the prediction of lower future productivity. In other words, high net-worth households are only good for an investment advisor’s productivity when the advisor is able to capture a large share of those households’ business, and, presumably, that household is less likely to leave.

COEFFICIENT PLOTS

Some authors have sought to describe interaction effects by producing tables of the coefficients, standard errors and confidence intervals for the focal independent variable with the moderator variable set to a finite number of values; these might be the mean and one standard deviation above and below the mean, or other values that meaningfully correspond to “low,” “medium” and “high” values of the moderator variable (Cohen et al. 2003; Jaccard and Turrisi 2003). This so-called “pick-a-point” approach (Rogosa 1980) has been criticized by proponents of more mathematically involved approaches for being somewhat arbitrary in the values of the moderator variable chosen, for choosing relatively few values in describing interaction effects, and also for failing to determine precisely the range of values of the moderator variable over which the focal independent variable exerts a significant effect (Bauer and Curran 2005; Hayes and Matthes 2009).

The approach I advocate here draws on the intuitiveness of the “pick-a-point” approach while automating the process of calculating coefficients and confidence intervals for the focal independent variable using many values of the moderator variable. Indeed, the macro capability of SAS® software makes this quite feasible. It is also motivated by the work of authors who have sought to use graphical techniques to depict the strength and direction of a moderated relationship while also representing the confidence intervals associated with model coefficients (Braumoeller 2004).

The logic of the program appearing in Fig. 13 is as follows. First, PROC STDIZE is used to mean-center all of the independent variables with the exception of the dichotomous TEAM variable and the two variables involved in the interaction effect, MEDIAN_HH_ASSETS and ACCTS_PER_HH. Next, the parmsint dataset that will contain the model coefficients, confidence intervals and the centering value of ACCTS_PER_HH is created in a DATA step (initially with zero observations).

The analytical heavy lifting is done in the next part of the program by the %INTPROBE macro. The macro performs the following steps iteratively with ACCTS_PER_HH centered between 1.0 and 6.0 in increments of 0.1. A DATA step is used to create the dataset data_3 containing the centered
Fig. 10

ODS OUTPUT BasicMeasures=means;
PROC UNIVARIATE DATA=data_1;
VAR ACCTS_PER_HH MEDIAN_HH_ASSETS TOTAL_REV_11;
RUN;

DATA _NULL_; 
SET means;
IF VarName="ACCTS_PER_HH" AND LocMeasure="Mean"
   THEN CALL SYMPUT('AVG_ACCTS_PER_HH', LocValue);
IF VarName="MEDIAN_HH_ASSETS" AND LocMeasure="Mean"
   THEN CALL SYMPUT('AVG_MEDIAN_HH_ASSETS', LocValue);
RUN;

DATA plot_1 (DROP=i j _MODEL_);
SET parmast (WHERE=(_MODEL_="INTERACTION") KEEP=_MODEL_ Intercept MEDIAN_HH_ASSETS ACCTS_PER_HH ACCTS_PER_HH_MED_HH_ASSETS RENAME=(MEDIAN_HH_ASSETS=b_MED_ASSETS ACCTS_PER_HH=b_ACCTS ACCTS_PER_HH_MED_HH_ASSETS=b_MED_ASSETS_ACCTS));
DO i=100 TO 600;
   DO j=1.5 TO 4 BY 0.5;
      MEDIAN_HH_ASSETS=i;
      MEDIAN_HH_ASSETS_CTR=i - INPUT(&AVG_MEDIAN_HH_ASSETS, BEST12.);
      ACCTS_PER_HH=j;
      ACCTS_PER_HH_CTR=j - INPUT(&AVG_ACCTS_PER_HH, BEST12.);
      PRED=
         Intercept + /* Intercept */
         (b_MED_ASSETS * MEDIAN_HH_ASSETS_CTR) + /* Median HH Assets */
         (b_ACCTS * ACCTS_PER_HH_CTR) + /* Accounts per household */
         (b_MED_ASSETS_ACCTS * (MEDIAN_HH_ASSETS_CTR * ACCTS_PER_HH_CTR)) /* Interaction */;
      OUTPUT;
   END;
END;
RUN;

DATA plot_2;
MERGE plot_1 (WHERE=(ACCTS_PER_HH=1.5) RENAME=(PRED=PRED_1_5)) plot_1 (WHERE=(ACCTS_PER_HH=2.0) RENAME=(PRED=PRED_2_0)) plot_1 (WHERE=(ACCTS_PER_HH=2.5) RENAME=(PRED=PRED_2_5)) plot_1 (WHERE=(ACCTS_PER_HH=3.0) RENAME=(PRED=PRED_3_0)) plot_1 (WHERE=(ACCTS_PER_HH=3.5) RENAME=(PRED=PRED_3_5)) plot_1 (WHERE=(ACCTS_PER_HH=4.0) RENAME=(PRED=PRED_4_0));
BY MEDIAN_HH_ASSETS;
RUN;
**Fig. 11**

ODS GRAPHICS ON /BORDER=OFF HEIGHT=2.5IN WIDTH=4IN;
ODS LISTING IMAGE_DPI=600 STYLE=JOURNAL SGE=OFF GPATH="C:\NESUG 2012";

PROC SGPLOT DATA=plot_2;
TITLE "Interactive Effect of Median HH Size and Accounts per HH";
SERIES Y=PRED_1_5 X=MEAN_HH_ASSETS
   /LINEATTRS=(THICKNESS=1 PATTERN=SOLID COLOR=CXE1E1E1)
   LEGENDLABEL="1.5 Accts per HH";
SERIES Y=PRED_2_0 X=MEAN_HH_ASSETS
   /LINEATTRS=(THICKNESS=1 PATTERN=SOLID COLOR=CXB4B4B4)
   LEGENDLABEL="2.0 Accts per HH";
SERIES Y=PRED_2_5 X=MEAN_HH_ASSETS
   /LINEATTRS=(THICKNESS=1 PATTERN=SOLID COLOR=CX878787)
   LEGENDLABEL="2.5 Accts per HH";
SERIES Y=PRED_3_0 X=MEAN_HH_ASSETS
   /LINEATTRS=(THICKNESS=1 PATTERN=SOLID COLOR=CX5A5A5A)
   LEGENDLABEL="3.0 Accts per HH";
SERIES Y=PRED_3_5 X=MEAN_HH_ASSETS
   /LINEATTRS=(THICKNESS=1 PATTERN=SOLID COLOR=CX2D2D2D)
   LEGENDLABEL="3.5 Accts per HH";
SERIES Y=PRED_4_0 X=MEAN_HH_ASSETS
   /LINEATTRS=(THICKNESS=1 PATTERN=SOLID COLOR=CO000000)
   LEGENDLABEL="4.0 Accts per HH";
KEYLEGEND /POSITION=RIGHT LOCATION=OUTSIDE ACROSS=1 DOWN=6 NOBORDER;
YAXIS MIN=400 MAX=1200 VALUES=(400 600 800 1000 1200) OFFSETMIN=0.02
   LABEL="Predicted Revenue (000s), 2011";
XAXIS MIN=100 MAX=600 VALUES=(100 200 300 400 500 600) OFFSETMIN=0.02
   LABEL="Median HH Size (000s)"
RUN;

ODS GRAPHICS OFF;

**Fig. 12**

Interactive Effect of Median HH Size and Accounts per HH

Predicted Revenue (000s), 2011

1.5 Accts per HH
2.0 Accts per HH
2.5 Accts per HH
3.0 Accts per HH
3.5 Accts per HH
4.0 Accts per HH
ACCTS_PER_HH variable and the product term ACCTS_PER_HH_MED_HH_ASSETS. Then, the regression model is run using PROC REG, and the regression coefficients output using the ODS OUTPUT statement are added to the parmsgn dataset using a series of intermediate DATA steps. Then, the datasets created within an iteration of the macro are deleted, and the next iteration of the macro begins with an incremented value for ACCTS_PER_HH. The final result is a dataset containing the necessary quantities to determine with a reasonable amount of precision the range of values of the moderator variable ACCTS_PER_HH over which the focal independent variable MEDIAN_HH_ASSETS has a significant effect on TOTAL_REV_11. These results further provide substantive interpretation of the effect of MEDIAN_HH_ASSETS on TOTAL_REV_11 when ACCTS_PER_HH is 1.0, 1.1, 1.2, and so on up to 6.0.

Partial output from the parmsgn dataset appears in Fig. 14. What this output confirms is that the effect of MEDIAN_HH_ASSETS on TOTAL_REV_11 is significant and negative when ACCTS_PER_HH is less than or equal to 2.2; its effect is significant and positive when ACCTS_PER_HH is greater than or equal to 3.3. When ACCTS_PER_HH is between 2.3 and 3.2, MEDIAN_HH_ASSETS has no significant effect on TOTAL_REV_11.

These results can also be summarized graphically by plotting the coefficients for MEDIAN_HH_ASSETS and their upper and lower confidence limits for selected values of ACCTS_PER_HH – for example, values 1.0 to 6.0 in increments of 0.5. A horizontal line at zero on the Y axis is helpful in determining whether confidence interval for MEDIAN_HH_ASSETS at a given value of ACCTS_PER_HH includes zero. (Fig. 15 provides the program for producing the coefficient plot, and the plot itself appears in Fig. 16.) This plot confirms what the above tabular output already indicated – that for ACCTS_PER_HH values of 2.5 and 3.0, MEDIAN_HH_ASSETS has no effect on TOTAL_REV_11. When ACCTS_PER_HH is at the low end of its range, the effect of MEDIAN_HH_ASSETS is significant and negative; when it is at the high end of its range, its effect is significant and positive.

**CONCLUSION**

Regression models containing interactive relationships certainly require greater care and nuance in their interpretation than simply additive models, but interpretable they remain. As I have shown in this paper, their specification and testing and their correct interpretation is made easier using different capabilities in SAS/STAT® and SAS/GRAPH®.

I should once more point out that the scope of this paper is limited to interaction effects in the context of linear regression. Further, the paper examined only a continuous-by-continuous (or bilinear) interaction effect. Of course, interaction effects can be implemented in a wide range of generalized linear models. They can also take the form of categorical-by-categorical and categorical-by-continuous interactions as befitting substantive theory and the data available to the researcher. Extending the concepts and tools discussed in this paper to these cases is relatively straightforward. I anticipate covering the specification, testing, probing and display of interaction effects in logistic regression and interactions involving more than one degree of freedom in a future paper.
Fig. 13

PROC STDIZE DATA=data_1 OUT=data_3 METHOD=MEAN;
VAR TRANS_REV_06 TRAILER_REV_06 FEE_REV_06 EXPERIENCE_YEARS
   CORE_HH_COUNT SMALL_HH_COUNT RETIREMENT_ACCT_COUNT;
RUN;

DATA parmsint;
LENGTH Variable $50 ACCTS_PER_HH_CENTER 3. Estimate LCL UCL 8.;
FORMAT p PVALUE6.4;
RUN;

%MACRO INTPROBE (DataIn=, DataOut= );
   %DO ACCT_HH=10 %TO 60;
      DATA &DataOut.;
      SET &DataIn.;
      CENTRE_VALUE=ROUND((&ACCT_HH*0.1), 0.1);
      ACCTS_PER_HH=ACCTS_PER_HH - CENTRE_VALUE ;
      ACCTS_PER_HH_MED_HH_ASSETS=ACCTS_PER_HH*MEDIAN_HH_ASSETS;
      RUN;
   proc reg data=&DataOut.;
      model total_rev_11=trans_rev_06 trailer_rev_06 fee_rev_06 experience_years team
         core_hh_count small_hh_count retirement_acct_count accts_per_hh
         median_hh_assets accts_per_hh_med_hh_assets
         /clb stb;
      ods output parameterestimates=parms;
      run; quit;
      data parmsstr (keep=variable estimate standardizedest lowercl uppercl probt
         accts_per_hh Center
         rename=(standardizedest=stdcoeff lowercl=lcl uppercl=ucl probt=p));
      length variable $50 accts_per_hh_center 3.;
      set parms;
      accts_per_hh_center=round((&acct_hh*0.1), 0.1); format accts_per_hh_center 3.1;
      run;
      data parmsint;
      set parmsint parmsstr;
      if variable="" then delete;
      format accts_per_hh_center 3.1;
      run;
   proc datasets lib=work nolist;
      delete &DataOut. parms parmsstr;
      run; quit;
   %end;
%MEND INTPROBE;

%INTPROBE(DataIn=data_3, DataOut=data_4);

PROC SORT DATA=parmsint;
BY Variable;
RUN;
### Fig. 14

**Effect of Median HH Assets at Different Values of Accounts per HH**

| Variable         | ACCTS_PER_HH_CENTER | Parameter | Lower 95% Parameter | Upper 95% Parameter | Pr > |t| |
|------------------|---------------------|-----------|---------------------|---------------------|------|---|
| MEDIAN_HH_ASSETS | 1.0                 | -0.915    | -1.462              | -0.368              | 0.0011 |
| MEDIAN_HH_ASSETS | 1.1                 | -0.859    | -1.374              | -0.344              | 0.0011 |
| MEDIAN_HH_ASSETS | 1.2                 | -0.803    | -1.286              | -0.319              | 0.0012 |
|                  |                     |           |                     |                     |      |   |
| MEDIAN_HH_ASSETS | 1.9                 | -0.409    | -0.697              | -0.121              | 0.0054 |
| MEDIAN_HH_ASSETS | 2.0                 | -0.353    | -0.619              | -0.086              | 0.0096 |
| MEDIAN_HH_ASSETS | 2.1                 | -0.296    | -0.545              | -0.048              | 0.0195 |
| MEDIAN_HH_ASSETS | 2.2                 | -0.240    | -0.475              | -0.006              | 0.0447 |
| MEDIAN_HH_ASSETS | 2.3                 | -0.184    | -0.409              | 0.041               | 0.1091 |
| MEDIAN_HH_ASSETS | 2.4                 | -0.128    | -0.349              | 0.093               | 0.2569 |
| MEDIAN_HH_ASSETS | 2.5                 | -0.072    | -0.294              | 0.151               | 0.5281 |
|                  |                     |           |                     |                     |      |   |
| MEDIAN_HH_ASSETS | 3.0                 | 0.210     | -0.090              | 0.509               | 0.1694 |
| MEDIAN_HH_ASSETS | 3.1                 | 0.266     | -0.058              | 0.590               | 0.1075 |
| MEDIAN_HH_ASSETS | 3.2                 | 0.322     | -0.028              | 0.673               | 0.0716 |
| MEDIAN_HH_ASSETS | 3.3                 | 0.378     | 0.000               | 0.757               | 0.0500 |
| MEDIAN_HH_ASSETS | 3.4                 | 0.435     | 0.027               | 0.842               | 0.0366 |
| MEDIAN_HH_ASSETS | 3.5                 | 0.491     | 0.054               | 0.928               | 0.0279 |
| MEDIAN_HH_ASSETS | 3.6                 | 0.547     | 0.079               | 1.015               | 0.0220 |
|                  |                     |           |                     |                     |      |   |

### Fig. 15

```
ODS GRAPHICS ON /BORDER=OFF HEIGHT=2.5IN WIDTH=4IN;
ODS LISTING IMAGE_DPI=600 STYLE=JOURNAL SGE=OFF GPATH="C:\NESUG 2012";

PROC SGPLOT DATA=parmsint (WHERE=(Variable="MEDIAN_HH_ASSETS"
   AND MOD(ACCTS_PER_HH_CENTER, 0.5)=0));
TITLE "Effect of Median HH Assets on Total Revenue, 2011";
SCATTER X=ACCTS_PER_HH_CENTER Y=Estimate
   /ERRORLOWER=LCL ERRORUPPER=UCL MARKERATTRS=(SYMBOL=CIRCLE)
   ERRORBARATTRS=(PATTERN=4);
XAXIS TYPE=DISCRETE OFFSETMIN=0.02 LABEL="Accounts per HH";
YAXIS MIN=-2 MAX=4 VALUES=(-2 TO 4 BY 1) LABEL="b (Median HH Assets)";
REFLINE 0 /AXIS=Y TRANSPARENCY=0.5;
RUN;

ODS GRAPHICS OFF;
```
REFERENCES


**ACKNOWLEDGEMENTS**

I would like to thank Patrick Kennedy for helpful discussions on developments in the retail wealth management industry and the interpretation of the regression results presented here. I would also like to thank Madeleine Cruickshank and Kieran Bol for their efforts in preparing the data analyzed in this paper.

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

**CONTACT INFORMATION**

Timothy B. Gravelle  
Principal Scientist & Director, Insights Lab  
PriceMetrix Inc.  
40 University Avenue, Suite 200  
Toronto, Ontario  
M5J 1T1  
CANADA  
Email: tim.gravelle@pricemetrix.com  
Web: [www.pricemetrix.com](http://www.pricemetrix.com)

PriceMetrix, the first choice in practice intelligence solutions for retail brokerages in North America, helps wealth management firms and their advisors, enhance revenue growth and increase profitability.

PriceMetrix  
REALIZE YOUR VALUE